future internet

Special Issue Reprint

Theory and Applications of Web 3.0 in the Media Sector

Edited by
Charalampos A. Dimoulas and Andreas Veglis

www.mdpi.com/journal/futureinternet

MDPI

Theory and Applications of Web 3.0 in the Media Sector

Theory and Applications of Web 3.0 in the Media Sector

Editors

Charalampos A. Dimoulas
Andreas Veglis

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Charalampos A. Dimoulas
Aristotle University of Thessaloniki
Greece

Andreas Veglis
Aristotle University of Thessaloniki
Greece

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Future Internet* (ISSN 1999-5903) (available at: https://www.mdpi.com/journal/futureinternet/special_issues/Media_Sector).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-7650-3 (Hbk)
ISBN 978-3-0365-7651-0 (PDF)

Contents

About the Editors

Charalampos A. Dimoulas

Dr. Charalampos A. Dimoulas was born in Munich, Germany, on 14 August 1974. He received his diploma and PhD from the School of Electrical and Computer Engineering of Aristotle University of Thessaloniki (AUTh) in 1997 and 2006, respectively. In 2008, he received a post-doctoral research scholarship on audiovisual processing and content management techniques for the intelligent analysis of prolonged multi-channel recordings at the Laboratory of Electronic Media (School of Journalism and Mass Communications, AUTh). He was elected lecturer (November 2009), assistant [rofessor (June 2014), associate professor (October 2018) and full professor (October 2022) of electronic media (audiovisual communication and media technologies) at the School of Journalism and Mass Communications, AUTh, where he is currently serving. Dr. Dimoulas has participated in over 30 national and international research projects and many respective scientific publications. His current scientific interests include media technologies, audiovisual signal processing, machine learning, multimedia semantics, cross-media authentication, digital audio and audiovisual forensics and more topics besides. Dr. Dimoulas is a member of IEEE, AES and the editorial board of the *Journal of the Audio Engineering Society* (JAES), contributing to discussions of signal processing and semantic audio (https://www.aes.org/journal/ed_board.cfm).

Andreas Veglis

Andreas Veglis is a professor of media technology, and head of the Media Informatics Lab at the School of Journalism and Mass Communication at the Aristotle University of Thessaloniki. He has served as an editor, member of scientific boards, and reviewer in various academic journals. Prof Veglis has more than 200 peer-reviewed papers on media technology and journalism. Specifically, he is the author or co-author of 12 books and 44 book chapters, published more than 100 papers in scientific journals and presented 144 papers at international and national conferences. Prof Veglis has been involved in 45 national and international research projects. His research interests include information technology in journalism, new media, algorithmic journalism, drone journalism, data journalism, big data, social media, open data, and content verification.

future internet

MDPI

Editorial

Theory and Applications of Web 3.0 in the Media Sector

Charalampos A. Dimoulas * and Andreas Veglis *

Multidisciplinary Media & Mediated Communication (M3C) Research Group, School of Journalism & Mass Communications, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
* Correspondence: babis@eng.auth.gr (C.A.D.); veglis@jour.auth.gr (A.V.)

Abstract: We live in a digital era, with vast technological advancements, which, among others, have a major impact on the media domain. More specifically, progress in the last two decades led to the end-to-end digitalization of the media industry, resulting in a rapidly evolving media landscape. In addition to news digitization, User-Generated Content (UGC) is dominant in this new environment, also fueled by Social Media, which has become commonplace for news publishing, propagation, consumption, and interactions. However, the exponential increase in produced and distributed content, with the multiplied growth in the number of plenary individuals involved in the processes, created urgent needs and challenges that need careful treatment. Hence, intelligent processing and automation incorporated into the Semantic Web vision, also known as Web 3.0, aim at providing sophisticated data documentation, retrieval, and management solutions to meet the demands of the new digital world. Specifically, for the sensitive news and media domains, necessities are created both at the production and consumption ends, dealing with content production and validation, as well as tools empowering and engaging audiences (professionals and end users). In this direction, state-of-the-art works studying news detection, modeling, generation, recommendation, evaluation, and utilization are included in the current Special Issue, enlightening multiple contemporary journalistic practices and media perspectives.

Keywords: web 3.0; semantic web; media industry; journalistic practices; journalism 3.0; news semantics; news recommendation; media automations; disinformation; hate speech

1. Introduction

In today's exploding Web landscape, vast amounts of information (documents, images, sounds, videos, multimedia storylines, etc.) are produced and published daily from various sources worldwide. As a result, the formation of the news agenda becomes tricky, as does the process of being credibly and reliably informed. Hence, plenary individuals, both in the roles of news consumers and content contributors (usually wearing the hat of citizen journalists), but also professional journalists and media communication experts, often find it difficult to retrieve specific and detailed information about a (complicated) topic to form a comprehensive informing or reporting view [1–4]. Today's news is mostly published in an irregular way, with multimedia assets (posts and articles, comments, reactions, etc.) propagating through a network of unstructured forms of data (and metadata). Users have to navigate multiple content instances and interconnecting nodes, lacking an efficient infrastructure to quickly discover, acquire, and analyze the information needed, which limits the news stream's exploitation prospects. Consequently, new challenges arise for algorithmic media automations concerning both news production and consumption ends (i.e., machine-assisted reporting, content selection/generation, validation, publishing, recommendation, retrieval, personalization, semantics, and so on) [5–11].

Since digital informing and mediated communication dominate today's ubiquitous society, content creation and publishing are no longer restricted to large organizations, with anyone being able to upload information in multiple formats (text, photos, audio, or video) [12]. Social media have emerged and broadly expanded to become the common

Citation: Dimoulas, C.A.; Veglis, A. Theory and Applications of Web 3.0 in the Media Sector. *Future Internet* **2023**, *15*, 165. https://doi.org/10.3390/fi15050165

Received: 24 April 2023
Accepted: 26 April 2023
Published: 28 April 2023

everyday practice for thousands of people, offering simplicity, immediacy, and interactivity [1–4,12]. Furthermore, the proliferation of mobile devices (smartphones, tablets, etc.) and their inherent content-capturing and networking capabilities fueled citizen and participatory journalism paradigms, enabling plenary users to contribute User-Generated Content (UGC) fast and effortlessly (i.e., personal or public information, such as news events/stories, articles, opinions, comments, etc.) [1,9–14]. At the same time, the plurality of data and news streams created urgent needs for better content documentation, validation, and management [1,4,9–15]. Hence, the advancement of the Semantic Web (SW) or Web 3.0 has been envisioned as a sophisticated solution to the above problems and challenges aspiring intelligent multimedia processing, analysis, and creation techniques [4,9,10,16].

Web 3.0 stands as the physical extension of the current Web where information is given a well-defined meaning, better enabling computers and people to work in cooperation [4,9–14]. Therefore, it can be seen as an Internet service with sophisticated technological features, in which proper/standardized documentation and semantic tagging will be delivered with the help of algorithms (in fully or semi-automated processes) [4,9,10]. These features attempt to form a Web environment in which humans and machines understand and interpret web-streamed information in the same context. The so-called SW services embody and integrate technologies aiming to complete the transition from today's Web of documents to a Web of well-documented data, where every piece of information will be accompanied by its semantic, conceptual, and contextual metadata. Fully structured and clarified relations to others (users, events, stories, models, etc.) will expedite data visualization and display purposes, also facilitating interoperability and integration between systems, environments, and applications, interconnecting concepts rather than just documents [4–10,16–19]. Thus, journalists and plenary individuals (content contributors and news consumers) will be able to efficiently discover, integrate, and reuse pieces of information from various sources. At this point, where the journalism and news industries intersect with Web 3.0, well-established journalistic practices/workflows are challenged by more sophisticated semantically enhanced procedures towards the transition to Journalism 3.0 [4–8,20,21]. As a result, the advanced technological framework unlocks various data exploitation capabilities, leading to higher functional levels. For instance, newsroom automations can expedite Web-Radio/-TV services with innumerous live streaming and post-processing augmentations [22,23], while they can also provide framing insights for better organizing blogs and news online [24] or data monitoring and gathering options for post-analysis purposes [25,26].

Contemporary Semantic Web and Big Data technologies are continuously elaborated to advance multimodal data analysis concerning information classification, semantic conceptualization and contextualization, content validation, and management automations, which can be primarily deployed in the sensitive news and media domains [1–4,16,19–29]. These modular automation and augmentation layers can fuel interaction mechanisms between corporations, machines, and individuals to accelerate crowdsourcing procedures for constructing and maintaining suitable media repositories [16,18,27,29]. Thereafter, digital literacy initiatives and audience engagement strategies can enhance the impact of SW services in the media world and the broader society. The current Special Issue enlightens the above perspectives through theoretical, algorithmic/technological, and case-study contributions, discussing challenges and state-of-the-art solutions on news detection, classification, (semantic) analysis and evaluation [12,18,27,29,30], automated modeling/prediction, generation, and recommendation of news content [19,31–33], evolution, aesthetics, and integration evaluation of Web 3.0 elements in news websites [34–36].

2. Contributions

The current Special Issue focused on enlightening the above multidisciplinary areas, inviting researchers to submit featured research works on intelligent media processes and practices related to news production, validation, management, publishing, etc., placing semantic services and metadata processing towards automated content generation,

recommendation, and assessment. Eleven (in total) contributions (*Ci*, $i = 1 \ldots 11$) were finally published within this Special Issue, elaborating on the perspectives of *Theory and Applications of Web 3.0 in the Media Sector*. The present section outlines the conducted works and their research outcomes, with Table 1 highlighting each article's scientific focus and contribution.

Table 1. Contributions by research areas, involved technologies, and proposed solutions.

Contributions	Research Area/Focus	Involved Technologies/Solutions
Contribution 1 (C1)	Hate speech detection and emotion analysis	Natural language processing and machine/deep learning algorithms in news semantics, web interface for data crowdsourcing and datasets creation
Contribution 2 (C2)	Financial forecasting from news websites and social networks using huge volumes of data (big data)	Natural language processing and machine/deep learning algorithms in stock forecasting, sentiment analysis, investment recommendations
Contribution 3 (C3)	Assessment of news homepages over the years (aesthetic-, functional-, content-wise)	Content analysis, aesthetics, and responsive design aspects (web, mobile, multiformat), audience analytics/insights, social instrument (interviews)
Contribution 4 (C4)	Automated news generation/publishing for earth observation/disaster reporting	Big data and artificial intelligence algorithms in breaking news detection, sentiment analysis, and automated/personalized news-report generation
Contribution 5 (C5)	Semantic preprocessing for breaking news detection (in Drone Journalism)	News streams monitoring/semantics, spatiotemporal and contextual detection of (breaking) news events, Drone Journalism recommendations
Contribution 6 (C6)	Semantic web integration analysis (in Art and Culture websites)	Contextual features, metrics, art/culture websites features, semantic web integration assessment
Contribution 7 (C7)	Audio semantic analysis and visualizations for audiovisual forensics	Semantic visualizations of machine/deep learning and signal processing to detect audio tampering
Contribution 8 (C8)	Social network analysis and visualization for tracing disinformation	Natural language processing and machine/deep learning algorithms in tracing and visualizing suspicious/inaccurate informatory streams
Contribution 9 (C9)	Analysis of aesthetic trends and semantic web adoption of media outlets	Automated archival data extraction and analysis to assess Semantic Web integration trends, DOM structure complexity, graphics, and color usage
Contribution 10 (C10)	News recommendation systems (NRS) modeling and evaluation	Experimental approach treating the NRS as a black box, entailing users as testers of algorithmic systems (algorithmic/collaborative audit methods)
Contribution 11 (C11)	Chatbot Media Automations in Professional Journalism	Experimental approach on the use of chatbot tools that are evaluated metric-wise and through social instruments (workshops)

The first paper presents the development and evaluation of a web interface (with its algorithmic backend) for creating and querying a multi-source database containing hate speech content [12]. Vrysis et al. (2021) implemented a Graphical User Interface (GUI) within the European project PHARM (Preventing Hate against Refugees and Migrants) to monitor and model hate speech against refugees and migrants in Greece, Italy, and Spain. The monitoring includes Social Media content, i.e., Twitter, YouTube, and Facebook comments and posts, as well as comments and articles from a selected list of websites, with the platform supporting the functionalities of searching (the formulated dataset), web-scrapping, and annotating additional records to contribute new samples to the repository. As an outcome, textual hate speech detection and sentiment analysis are provided using novel methods and machine learning algorithms, which can be used either for tracking and evaluating external web streams or for self-checking articles before making them public, also supporting media literacy. The interface and the involved methods are objectively (metric-based) and subjectively assessed, with the gained positive evaluation confirming the approach's usefulness and the interface's usability (Contribution 1).

The second paper focuses on automated stock forecasting using both financial and textual data from news websites and social networks (Twitter, Stocktwits), combining methods from various scientific fields, such as information retrieval, natural language

processing, and deep learning [31]. Theodorou et al. (2021) present the supportive platform ASPENDYS, developed as part of the homonymous European research project, intending to facilitate the management and decision making of investment actions through personalized recommendations. The implicated processing relies on technical analysis and machine learning methods for the financial data treatment, with textual data being analyzed in terms of reliability and sentiments towards an investment. As an outcome, investment signals are generated for a certain transaction combining the financial and sentiment analysis insights, which are finally recommended to the investors. A watchful assessment is conducted concerning the interface and its functionalities (i.e., portfolio management, sentiment analysis, extracted investment signals), with the application use cases illustrating practical uses and validating the approach's helpfulness and impact (Contribution 2).

The third paper deals with the evolution of news presentation in online newspapers, monitoring their visual progress from simple digital editions that merely served to dump content from print newspapers (rigidity) to sophisticated multi-format multimedia products with interactive features (exuberance) [34]. Peña Fernández, Casado del Río, and García-González (2021) conducted a longitudinal study on the design of online media, analyzing the front pages of five general information Spanish newspapers (elpais.com, elmundo.es, abc.es, lavanguardia.com, and elperiodico.com (accessed on 20 April 2023)) over the past 25 years (1996–2020). Further, six interviews were conducted in parallel with managers of different online media outlets. The evaluation results, combining content analysis and subjective assessment of the interviewees' responses, revealed an evolution from static and rigid layouts to dynamic, mobile, and responsive formats, displaying a balance between text and visual elements. The analysis included the language used, multimedia features, audience habits, and the degree of the offered interactions. Hence, without explicitly tackling semantic services evolution, the current work indicated presentation and functional changes in the online media frontpages, some of which are triggered by shifting to Web 3.0, while others point to the need for further semantic automations, customizations, and personalization in the upcoming Web eras, 3.0 and beyond (Contribution 3).

The fourth paper focuses on collecting and processing diverse and heterogeneous information, where multimedia data can be extracted from different sources on the Web [19]. In the context of the Journalism 3.0 vision, Tzouma, Zamichos, Efthymiadis, Drosou, and Tzovaras (2021) explore the possibility of creating a tool for utilizing Earth observations, i.e., to manage the massive volumes of image data, thus helping media industry professionals in the selection, usage, and dissemination of such (news) content to the public. Hence, intending to make productive satellite images for professionals who are not familiar with image processing (as other related tools require), a novel platform is implemented to automate some of the journalistic practices, i.e., to detect and receive breaking news information early in real time (especially for events related to disasters, such as floods and fires) to retrieve and collect Earth observation images for a certain event and to automatically compose personalized articles adapted to the authors' writing styles. The crafted EarthPress platform comprises the user interface, the user manager, the database manager, the breaking news detector, the data fusion, the data/image processing, the journalist profile extractor, and the software bot (EarthBot) that is responsible for the text synthesis, thus containing dominant semantic web features. Based on the conducted analysis and assessment, EarthPress represents an added-value tool, not only for professional journalists or editors in the media industry but also for freelancers and article writers who use the extracted information and data in their articles (Contribution 4).

The fifth paper casts light on the semantic preprocessing of Web and Social Media informatory streams, aiming to detect breaking news events, especially those suited for drone coverage (e.g., physical disasters, such as earthquakes or storms, fire or traffic accidents, traffic jam problems, etc.) [18]. Niarchos, Stamatiadou, Dimoulas, Veglis, and Symeonidis (2021) elaborate on the need for news validation and documentation using piece-of-evidence material, such as visual and multimedia documents of photo/video footage. While reporters and mobile journalists can serve this requirement, a quick on-site

presence is not always feasible due to access or distance/time difficulties that might cause unwanted delays and poor capturing quality. To face these demands, Drone Journalism (DJ) uses Unmanned Aerial Vehicles (UAVs)/drones to help journalists and news organizations capture and share breaking news stories. The current paper envisions a DJ framework to mediate real-time breaking news coverage, introducing a data retrieval and semantics preprocessing approach to detect and classify news events suitable for DJ coverage. Based on this, breaking news alerts/notifications are sent to drone operators along with automated preparations of flight plans, embodying the existing regulatory framework, security, and ethical matters. Backend implementation and pilot evaluation of the proposed system are conducted, with a modular architecture facilitating the integration of news alerts sent by mobile devices, elaborating on the inherent localization and networking capabilities to extract time-, location-, and context-aware semantic metadata. The pilot results and the received feedback rated the proposed approach useful in providing the contextual and spatiotemporal attributes of breaking news, with a more holistic coverage of the events offered by combining diverse drone footage and UGC mobile streams (Contribution 5).

The sixth paper addresses the fields of art and culture, some of the most eager to integrate with the Semantic Web since metadata, data structures, linked (open) data, and other building blocks of this Web of Things are considered essential in cataloging and disseminating art- and culture-related content (e.g., the Getty vocabularies project and the Europeana initiative) [35]. Giannakoulopoulos et al. (2022), motivated by the constantly evolving nature of art, which is the subject of many journalist blogs and websites, proceeded to investigate the use of Semantic Web technologies in media outlets that diffuse art- and culture-related content. The study formulates quantitative metrics to evaluate Semantic Web integration in art and culture media outlets, analyzing the impact of that integration on websites' popularity in the modern competitive landscape of the Web. A vast array of art-related media outlets was investigated, ranging greatly in size and popularity, based on a variety of metrics that were consolidated into a comprehensive integration rating. Consequently, the connection between Semantic Web integration and popularity was analyzed through a gradient boosting analysis. They conclude that studying and analyzing the tangible presence of the Semantic Web are vital steps to monitor its progress and stay on course to achieve its true potential, which, so far, remains largely untapped. Apart from its importance in art and culture media, the conducted research methodology and practical implementation may be extended to multiple topics/domains and broader multidisciplinary collaborations in the news and media industries, where semantic services are expected to have a highly positive impact (Contribution 6).

The seventh paper focuses on the development of a computer-supported toolbox with online functionality for assisting technically inexperienced users (journalists or the public) in visually investigating the consistency of audio streams to detect potential interventions coupled with disinformation [29]. Vryzas, Katsaounidou, Vrysis, Kotsakis, and Dimoulas (2022) elaborated on previous research [37,38] to set an audio forensics web environment (which is very limited), emanating on the photo/image forensics examples (and their offered functionalities), with multiple related platforms being already available online [39]. The proposed framework incorporates several algorithms on its backend implementation, including a novel CNN model that performs a Signal-to-Reverberation ratio (SRR) estimation with a mean square error of 2.9%. Hence, it is, for instance, feasible to monitor the conditions of the sound-capturing site (i.e., the "room acoustics") to detect recording inconsistencies and possible audio tampering. Users can access the application online to upload the audio/video file (or YouTube link) they want to inspect audio-wise. Then, a set of interactive visualizations are generated as outcomes of Digital Signal Processing and Machine Learning models, facilitating audio continuity and consistency evaluation. Users can evaluate the authenticity of the dataset samples, with files stored in the database supplemented by analysis results and crowdsourced annotations. Audio semantics bring added value to audiovisual forensics and multimedia disinformation detection, featuring lighter processing (compared to video) with the sound's inherent continuity (i.e., the sound

is always present in a recording, regardless of the microphone steering and the camera's viewing angle). Pilot evaluation results validated the usefulness of the aimed functionality, also considering that very few related applications exist (Contribution 7).

The eighth paper also focuses on the online misinformation problem, introducing a tool for analyzing the social web and gaining insights into communities that drive misinformation online [27]. More specifically, Papadopoulou et al. (2022) present the MeVer NetworkX analysis and visualization tool, which helps users delve into Social Media conversations, gaining insights about how information propagates and accumulating intuition about communities formed via interactions. The multidisciplinary contribution of MeVer lies in its easy navigation through a multitude of features, providing valuable insights about the account behaviors and data propagation in Online Social Networks, i.e., Twitter, Facebook, and Telegram graphs, while also encompassing the modularity to integrate more platforms. Four Twitter datasets related to COVID-19 disinformation were utilized to present the tool's functionalities and evaluate its effectiveness. As the authors conclude, to the best of their knowledge, MeVer stands as the only tool supporting the analysis of multiple platforms and even providing cross-platform investigations, aiming at facilitating the demanding work of journalists and fact checkers to combat disinformation. The presented use cases utilizing the advanced functionalities offered by the tool validated the usefulness and impact of the approach. For instance, aggregation and visualization capabilities provide easy ways to navigate large graphs without special knowledge. Hence, the crafted functionalities usher in semi-automatic procedures that increase productivity, promote cooperation, and save time, making the tool applicable even to average users (Contribution 8).

The ninth paper emphasizes aesthetic trends and Semantic Web adoption of media outlets, as identified through automated archival data extraction and analysis processes [36]. Lamprogeorgos, Pergantis, Panagopoulos, and Giannakoulopoulos (2022) employed various web data extraction techniques to collect current and archival information from popular news websites in Greece to monitor and record their progress through time. Specifically, HTML source code and homepage screenshots were collected for a large number of websites (the top 1000 online media outlets based on Web traffic) using automated archival data extraction techniques to investigate the evolution of their homepage throughout different time instances for two decades. This gathered information was used to identify Semantic Web integration trends, Document Object Model (DOM) structure complexity, number of graphics, color usage, and more. The identified trends were analyzed and discussed as a means to gain a better understanding of the ever-changing presence of the media industry on the Web, with the evolution of Semantic Web technologies proving to be rapid and extensive in online media outlets. Furthermore, website structural and visual complexity presented a steady and significant positive trend, accompanied by increased adherence to color harmony. In conclusion, the study underlines the constantly evolving World Wide Web, influenced both by the rise and fall of technologies and by the continuous changes in human nature through cultural trends, global events, and globalization in general. The conducted study and its novel methods can be extended to provide valuable knowledge pertaining not only to the present but hopefully preparing us for the future. In the end, tracing the advancements of the Semantic Web and the aesthetic evolution of user interfaces can be valuable tools at the disposal of every online media outlet (Contribution 9).

The tenth paper deploys an experimental approach to model and validate a news recommending system (NRS) in a mainstream medium-sized news organization [32]. Spyridou, Djouvas, and Milioni (2022) examined the performance of a ready-to-use (of the shelf) NRS application by observing its outputs. Specifically, using an experimental design entailing users as system testers, the authors analyzed the composition of the personalized MyNews area on the basis of accuracy and user engagement. Addressing the development of algorithms for news media that differ from other media offerings in terms of their civic role, a two-fold aim was pursued: first, to identify the implicated parameters and discover the underlying algorithmic functionality, and second, to evaluate, in practice,

the NRS efficiency through the deployed experimentation. Results indicate that while the algorithm manages to adapt between different users based on their past behavior, overall it underperforms due to flawed design decisions rather than technical deficiencies. The requirement to populate the personalized agenda with a large number of news items, the imperative of recency, the problem of unsystematic tagging and the underuse of available content reduced the capacity of the algorithm to offer a successful personalized agenda. As an outcome, the study offers insights to guide/improve NRS design, considering the production capabilities of the news outlets while supporting their business goals along with users' demands and journalism's civic values. Despite not being a core technological work, this research offers valuable feedback for developing and implementing content recommendation algorithmic solutions for news offering (Contribution 10).

The eleventh paper approaches the current issue from the perspective of interactivity and chatbots, which started infiltrating the media sphere [33]. More precisely, Kotenidis, Vryzas, Veglis, and Dimoulas (2022) focused on new/innovative ways offered by chatbots to news outlets in creating and sharing their content, with an even larger emphasis on back-and-forth communication and news-reporting personalization. The research highlights two important factors to assess the integration efficiency of chatbots in professional journalism. Firstly, the chatbot programming feasibility by plenary individuals without technological background (journalists and media professionals in the current scenario) using low-code platforms. Secondly, the usability of the crafted chatbot news-reporting agents, as perceived by the targeted audience (broader news consumers). Hence, today's most popular chatbot creation platforms are analyzed and assessed within a three-phase evaluation framework. First, the offered interactivity features are evaluated within an appropriate metrics framework. Second, a two-part workshop is conducted with journalists operating the selected platforms (with minimum training) to create their chatbot agents for news reporting. Third, the crafted chatbots are evaluated by a larger audience concerning the usability and overall user experience. The study found that all three platforms received positive evaluations, with high usefulness and usability scores. Professional journalists expressed their confidence in using the suggested platforms for chatbot design, which implies an important attitude change, given that attendees were either unaware or skeptical before the experimental process and the quick guiding. Thus, chatbots are, in fact, suitable for achieving some of the wanted media automations, without requiring prior knowledge of semantic web technologies (Contribution 11).

Based on the provided insights and inspired by [16], an overview of the works included in the Special Issue is presented in Figure 1, mapping their role and contribution across a generic end-to-end model of semantic media services and automations. The discussed topics and the associated solutions are depicted with their future extensions, thus forming a holistic vision encompassing important milestones of adopting Web 3.0 technology (and beyond) in the media industry. Among others, the diagram projects the relation and complementarity of the eleven scientific contributions (*Ci*) to the different model phases and functionalities. Hence, the research works included in this Special Issue are highly representative, appropriately demonstrating the main processes of the end-to-end chain. Nevertheless, future multidisciplinary research and collaborations are also highlighted and anticipated, augmenting the outcomes and the impact of the current "Future Internet Special Issue *Theory and Applications of Web 3.0 in the Media Sector*".

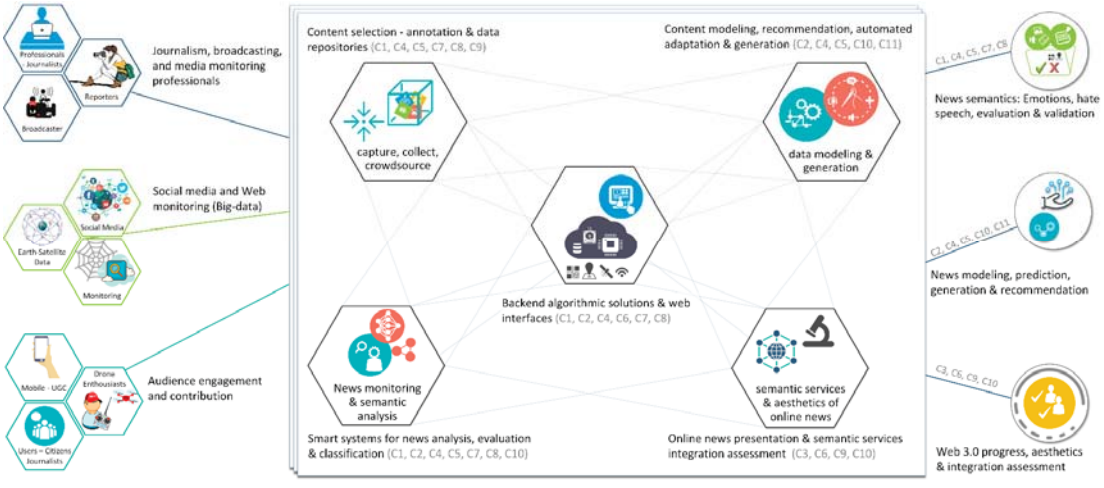


Figure 1. Future Internet volume on *Theory and Applications of Web 3.0 in the Media Sector*: a generic end-to-end model embodying the discussed topics/solutions and a generic one [16].

Table 2 lists all eleven (11) contributions incorporated in this Special Issue with their associated citations.

Table 2. List of contributions with their associated citations.

Contribution 1 (C1) [12]	Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. https://doi.org/10.3390/fi13030080
Contribution 2 (C2) [31]	Theodorou, T.-I.; Zamichos, A.; Skoumperdis, M.; Kougioumtzidou, A.; Tsolaki, K.; Papadopoulos, D.; Patsios, T.; Papanikolaou, G.; Konstantinidis, A.; Drosou, A.; Tzovaras, D. An AI-Enabled Stock Prediction Platform Combining News and Social Sensing with Financial Statements. *Future Internet* **2021**, *13*, 138. https://doi.org/10.3390/fi13060138
Contribution 3 (C3) [34]	Peña-Fernández, S.; Casado-del-Río, M.Á.; García-González, D. From Rigidity to Exuberance: Evolution of News on Online Newspaper Homepages. *Future Internet* **2021**, *13*, 150. https://doi.org/10.3390/fi13060150
Contribution 4 (C4) [19]	Tsourma, M.; Zamichos, A.; Efthymiadis, E.; Drosou, A.; Tzovaras, D. An AI-Enabled Framework for Real-Time Generation of News Articles Based on Big EO Data for Disaster Reporting. *Future Internet* **2021**, *13*, 161. https://doi.org/10.3390/fi13060161
Contribution 5 (C5) [18]	Niarchos, M.; Stamatiadou, M.E.; Dimoulas, C.; Veglis, A.; Symeonidis, A. A Semantic Preprocessing Framework for Breaking News Detection to Support Future Drone Journalism Services. *Future Internet* **2022**, *14*, 26. https://doi.org/10.3390/fi14010026
Contribution 6 (C6) [35]	Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Kouretsis, A.; Lamprogeorgos, A.; Varlamis, I. Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets. *Future Internet* **2022**, *14*, 36. https://doi.org/10.3390/fi14020036
Contribution 7 (C7) [29]	Vryzas, N.; Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing. *Future Internet* **2022**, *14*, 75. https://doi.org/10.3390/fi14030075
Contribution 8 (C8) [27]	Papadopoulou O., Makedas T., Apostolidis L., Poldi F., Papadopoulos S., Kompatsiaris I. MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation. *Future Internet*. 2022; 14(5):147. https://doi.org/10.3390/fi14050147
Contribution 9 (C9) [36]	Lamprogeorgos, A.; Pergantis, M.; Panagopoulos, M.; Giannakoulopoulos, A. Aesthetic Trends and Semantic Web Adoption of Media Outlets Identified through Automated Archival Data Extraction. *Future Internet* **2022**, *14*, 204. https://doi.org/10.3390/fi14070204
Contribution 10 (C10) [32]	Spyridou, P.; Djouvas, C.; Milioni, D. Modeling and Validating a News Recommender Algorithm in a Mainstream Medium-Sized News Organization: An Experimental Approach. *Future Internet* **2022**, *14*, 284. https://doi.org/10.3390/fi14100284
Contribution 11 (C11) [33]	Kotenidis, E.; Vryzas, N.; Veglis, A.; Dimoulas, C. Integrating Chatbot Media Automations in Professional Journalism: An Evaluation Framework. *Future Internet* **2022**, *14*, 343. https://doi.org/10.3390/fi14110343

3. Conclusions

With the complete digitalization of the end-to-end media processes, the interest has shifted to automating content production, distribution, and management. The so-called Semantic Web services are already present in the media industry, facilitating the works of both professional journalists and the broader news-consuming audience through the offered functionalities of automatic news/data generation, adaptation/personalization, recommendation, and retrieval. Representative works published in this Special Issue verified that notable progress has been made, with significant ongoing multidisciplinary research focusing on the domains of journalism and media, encompassing multiple angles (technological, algorithmic, journalistic, communicational, social, pedagogical, etc.). Nevertheless, further research needs to be conducted for the transition to the new media environment to be completed.

In today's highly diversified and ubiquitous society, where vast volumes of data and informatory streams are uncontrollably distributed among multiple networking terminals, users, and communities, critical processes of data evaluation and management need to be supported by technological means and addressed through interdisciplinary approaches. The portrayed conclusions also line up that people and broader society should not be defensive towards upcoming technologies and services that they are currently unaware and skeptical of but, instead, should be willing to become actively involved in the conducted evolutions from which they can only earn knowledge, skills, and digital literacy. Previous experience has shown whatever (media) tools are helpful to the targeted users will prevail in the end, no matter what, with the critical question shifting to how quickly and efficiently an optimal and fair configuration can be reached. Hence, a cooperative spirit among multiple disciplines is necessary to most appropriately shape these new trends to benefit our societies and democracies, i.e., serving the citizens' rights for objective, timely, and reliable news informing, also aligning with the broader civic values of journalism.

Data Availability Statement: Data supporting this article can be found in the listed contributions and their associated Data Availability Statements.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018.
2. Siapera, E.; Veglis, A. *The Handbook of Global Online Journalism*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Saridou, T.; Veglis, A. Exploring the Integration of User-Generated Content in Media Organizations through Participatory Journalism. In *Encyclopedia of Information Science and Technology*, 5th ed.; IGI Global: Hershey, PA, USA, 2021; pp. 1152–1163.
4. Matsiola, M.; Dimoulas, C.A.; Kalliris, G.; Veglis, A.A. Augmenting User Interaction Experience through Embedded Multimodal Media Agents in Social Networks. In *Information Retrieval and Management*; IGI Global: Hershey, PA, USA, 2018; pp. 1972–1993.
5. Diakopoulos, N. *Automating the News: How Algorithms are Rewriting the Media*; Harvard University Press: Harvard, UK, 2019.
6. Diakopoulos, N. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digit. J.* **2020**, *8*, 945–967. [CrossRef]
7. Thurman, N.; Lewis, S.; Kunert, J. Algorithms, Automation, and News. *Digit. J.* **2019**, *7*, 980–992. [CrossRef]
8. Thurman, N.; Schifferes, S. The Future of Personalization at News Websites. *J. Stud.* **2012**, *13*, 775–790. [CrossRef]
9. Dimoulas, C.A.; Veglis, A.A.; Kalliris, G.; Khosrow-Pour, D.M. Semantically Enhanced Authoring of Shared Media. In *Encyclopedia of Information Science and Technology*, 4th ed.; IGI Global: Hershey, PA, USA, 2018; pp. 6476–6487.
10. Saridou, T.; Veglis, A.; Tsipas, N.; Panagiotidis, K. Towards a Semantic-Oriented Model of Participatory Journalism Management. Available online: https://coming.gr/wp-content/uploads/2020/02/2_2019_JEICOM_SPissue_Saridou_pp.-27-37.pdf (accessed on 18 March 2021).
11. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. Machine-assisted reporting in the era of Mobile Journalism: The MOJO-mate platform. *Strategy Dev. Rev.* **2019**, *9*, 22–43.
12. Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. [CrossRef]
13. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How in Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173. [CrossRef]

14. Stamatiadou, M.E.; Thoidis, I.; Vryzas, N.; Vrysis, L.; Dimoulas, C. Semantic Crowdsourcing of Soundscapes Heritage: A Mojo Model for Data-Driven Storytelling. *Sustainability* **2021**, *13*, 2714. [CrossRef]
15. Cammaerts, B. Radical pluralism and free speech in online public spaces. *Int. J. Cult. Stud.* **2009**, *12*, 555–575. [CrossRef]
16. Dimoulas, C.A. Cultural Heritage Storytelling, Engagement and Management in the Era of Big Data and the Semantic Web. *Sustainability* **2022**, *14*, 812. [CrossRef]
17. Pileggi, S.F.; Fernandez-Llatas, C.; Traver, V. When the Social Meets the Semantic: Social Semantic Web or Web 2.5. *Future Internet* **2012**, *4*, 852–864. [CrossRef]
18. Niarchos, M.; Stamatiadou, M.E.; Dimoulas, C.; Veglis, A.; Symeonidis, A. A Semantic Preprocessing Framework for Breaking News Detection to Support Future Drone Journalism Services. *Future Internet* **2022**, *14*, 26. [CrossRef]
19. Tsourma, M.; Zamichos, A.; Efthymiadis, E.; Drosou, A.; Tzovaras, D. An AI-Enabled Framework for Real-Time Generation of News Articles Based on Big EO Data for Disaster Reporting. *Future Internet* **2021**, *13*, 161. [CrossRef]
20. Dörr, K.N. Mapping the field of Algorithmic Journalism. *Digit. J.* **2015**, *4*, 700–722. [CrossRef]
21. Panagiotidis, K.; Veglis, A. Transitions in Journalism—Toward a Semantic-Oriented Technological Framework. *J. Media* **2020**, *1*, 1. [CrossRef]
22. Vryzas, N.; Vrysis, L.; Dimoulas, C. Audiovisual speaker indexing for Web-TV automations. *Expert Syst. Appl.* **2022**, *186*, 115833. [CrossRef]
23. Vryzas, N.; Tsipas, N.; Dimoulas, C. Web Radio Automation for Audio Stream Management in the Era of Big Data. *Information* **2020**, *11*, 205. [CrossRef]
24. Touri, M.; Kostarella, I. News blogs versus mainstream media: Measuring the gap through a frame analysis of Greek blogs. *Journalism* **2017**, *18*, 1206–1224. [CrossRef]
25. Kostarella, I.; Kotsakis, R. The Effects of the COVID-19 "Infodemic" on Journalistic Content and News Feed in Online and Offline Communication Spaces. *J. Media* **2022**, *3*, 471–490. [CrossRef]
26. Tsipas, N.; Vrysis, L.; Konstantoudakis, K.; Dimoulas, C. Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings. *J. Acoust. Soc. Am.* **2020**, *148*, 3751–3761. [CrossRef]
27. Papadopoulou, O.; Makedas, T.; Apostolidis, L.; Poldi, F.; Papadopoulos, S.; Kompatsiaris, I. MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation. *Future Internet* **2022**, *14*, 147. [CrossRef]
28. Veglis, A.; Saridou, T.; Panagiotidis, K.; Karypidou, C.; Kotenidis, E. Applications of Big Data in Media Organizations. *Soc. Sci.* **2022**, *11*, 414. [CrossRef]
29. Vryzas, N.; Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing. *Future Internet* **2022**, *14*, 75. [CrossRef]
30. Kotenidis, E.; Veglis, A. Algorithmic Journalism—Current Applications and Future Perspectives. *J. Media* **2021**, *2*, 244–257. [CrossRef]
31. Theodorou, T.-I.; Zamichos, A.; Skoumperdis, M.; Kougioumtzidou, A.; Tsolaki, K.; Papadopoulos, D.; Patsios, T.; Papanikolaou, G.; Konstantinidis, A.; Drosou, A.; et al. An AI-Enabled Stock Prediction Platform Combining News and Social Sensing with Financial Statements. *Future Internet* **2021**, *13*, 138. [CrossRef]
32. Spyridou, P.; Djouvas, C.; Milioni, D. Modeling and Validating a News Recommender Algorithm in a Mainstream Medium-Sized News Organization: An Experimental Approach. *Future Internet* **2022**, *14*, 284. [CrossRef]
33. Kotenidis, E.; Vryzas, N.; Veglis, A.; Dimoulas, C. Integrating Chatbot Media Automations in Professional Journalism: An Evaluation Framework. *Future Internet* **2022**, *14*, 343. [CrossRef]
34. Peña-Fernández, S.; Casado-del-Río, M.Á.; García-González, D. From Rigidity to Exuberance: Evolution of News on Online Newspaper Homepages. *Future Internet* **2021**, *13*, 150. [CrossRef]
35. Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Kouretsis, A.; Lamprogeorgos, A.; Varlamis, I. Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets. *Future Internet* **2022**, *14*, 36. [CrossRef]
36. Lamprogeorgos, A.; Pergantis, M.; Panagopoulos, M.; Giannakoulopoulos, A. Aesthetic Trends and Semantic Web Adoption of Media Outlets Identified through Automated Archival Data Extraction. *Future Internet* **2022**, *14*, 204. [CrossRef]
37. Vryzas, N.; Katsaounidou, A.; Kotsakis, R.; Dimoulas, C.A.; Kalliris, G. Investigation of audio tampering in broadcast content. In Proceedings of the Audio Engineering Society Convention 144, Milan, Italy, 23–26 May 2018.
38. Vryzas, N.; Katsaounidou, A.; Kotsakis, R.; Dimoulas, C.A.; Kalliris, G. Audio-driven multimedia content authentication as a service. In Proceedings of the Audio Engineering Society Convention 146, Dublin, Ireland, 20–23 March 2019.
39. Katsaounidou, A.; Gardikiotis, A.; Tsipas, N.; Dimoulas, C. News authentication and tampered images: Evaluating the photo-truth impact through image verification algorithms. *Heliyon* **2020**, *6*, e05808. [CrossRef]

future internet

MDPI

Article

A Web Interface for Analyzing Hate Speech

Lazaros Vrysis [1], Nikolaos Vryzas [1], Rigas Kotsakis [1], Theodora Saridou [1], Maria Matsiola [1], Andreas Veglis [1,*], Carlos Arcila-Calderón [2] and Charalampos Dimoulas [1]

[1] School of Journalism & Mass Communication, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; lvrysis@auth.gr (L.V.); nvryzas@auth.gr (N.V.); rkotsakis@auth.gr (R.K.); saridout@jour.auth.gr (T.S.); mmat@jour.auth.gr (M.M.); babis@eng.auth.gr (C.D.)

[2] Facultad de Ciencias Sociales, Campus Unamuno, University of Salamanca, 37007 Salamanca, Spain; carcila@usal.es

* Correspondence: veglis@jour.auth.gr

Abstract: Social media services make it possible for an increasing number of people to express their opinion publicly. In this context, large amounts of hateful comments are published daily. The PHARM project aims at monitoring and modeling hate speech against refugees and migrants in Greece, Italy, and Spain. In this direction, a web interface for the creation and the query of a multi-source database containing hate speech-related content is implemented and evaluated. The selected sources include Twitter, YouTube, and Facebook comments and posts, as well as comments and articles from a selected list of websites. The interface allows users to search in the existing database, scrape social media using keywords, annotate records through a dedicated platform and contribute new content to the database. Furthermore, the functionality for hate speech detection and sentiment analysis of texts is provided, making use of novel methods and machine learning models. The interface can be accessed online with a graphical user interface compatible with modern internet browsers. For the evaluation of the interface, a multifactor questionnaire was formulated, targeting to record the users' opinions about the web interface and the corresponding functionality.

Keywords: hate speech detection; natural language processing; web interface; database; machine learning; lexicon; sentiment analysis; news semantics

Citation: Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. https://doi.org/10.3390/fi13030080

Academic Editor: Devis Bianchini

Received: 26 February 2021
Accepted: 18 March 2021
Published: 22 March 2021

1. Introduction

In today's ubiquitous society, we experience a situation where digital informing and mediated communication are dominant. The contemporary online media landscape consists of the web forms of the traditional media along with new online native ones and social networks. Content generation and transmission are no longer restricted to large organizations and anyone who wishes may frequently upload information in multiple formats (text, photos, audio, or video) which can be updated just as simple. Especially, regarding social media, which since their emergence have experienced a vast expansion and are registered as an everyday common practice for thousands of people, the ease of use along with the immediacy they present made them extremely popular. In any of their modes, such as microblogging (like Twitter), photos oriented (like Instagram), etc., they are largely accepted as fast forms of communication and news dissemination through a variety of devices. The portability and the multi-modality of the equipment employed (mobile phones, tablets, etc.), enables users to share, fast and effortless, personal or public information, their status, and opinions via the social networks. Thus, communications nodes that serve many people have been created minimizing distances and allowing free speech without borders; since more voices are empowered and shared, this could serve as a privilege to societies [1–8]. However, in an area that is so wide and easily accessible to large audiences many improper intentions with damaging effects might be met as well, one of which is hate speech.

It is widely acknowledged that xenophobia, racism, gender issues, sexual orientation, and religion among others are topics that trigger hate speech. Although no universally agreed definition of hate speech has been identified, the discussion originates from discussions on freedom of expression, which is considered one of the cornerstones of a democracy [9]. According to Fortuna and Nunes (2018, p. 5): "Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used" [10]. Although international legislation and regulatory policies based on respect for human beings prohibit inappropriate rhetoric, it finds ways to move into the mainstream, jeopardizing values that are needed for societal coherence and in some cases relationships between nations, since hate speech may fuel tensions and incite violence. It can be met towards one person, a group of persons, or to nobody in particular [11] making it a hard to define and multi-dimensional problem. Specifically, in Europe as part of the global North, hate speech is permeating public discourse particularly subsequent to the refugee crisis, which mainly -but not only- was ignited around 2015 [12]. In this vein, its real-life consequences are also growing since it can be a precursor and incentive for hate crimes [13].

Societal stereotypes enhance hate speech, which is encountered both in real life and online, a space where discourses are initiated lately around the provision of free speech without rules that in some cases result to uncontrolled hate speech through digital technologies. Civil society apprehensions led to international conventions on the subject and even further social networking sites have developed their own services to detect and prohibit such types of expressed rhetoric [14], which despite the platforms' official policies as stated in their terms of service, are either covert or overt [15]. Of course, a distinction between hate and offensive speech must be set clear and this process is assisted by the definition of legal terminology. Mechanisms that monitor and further analyze abusive language are set in efforts to recognize aggressive speech expanding on online media, to a degree permitted by their technological affordances. The diffusion of hateful sentiments has intrigued many researchers that investigate online content [11,13,15,16] initially to assist in monitoring the issue and after the conducted analysis on the results, to be further promoted to policy and decision-makers, to comprehend it in a contextualized framework and seek for solutions.

Paz, Montero-Díaz, and Moreno-Delgado (2020, p. 8) refer to four factors, media used to diffuse hate speech, the subject of the discourse, the sphere in which the discourse takes place, and the roots or novelty of the phenomenon and its evolution that each one offers quantification and qualification variables which should be further exploited through diverse methodologies and interdisciplinarity [17]. In another context, anthropological approaches and examination of identities seek for the genealogy through which hate speech has been created and sequentially moved to digital media as well as the creation of a situated understanding of the communication practices that have been covered by hate speech [13]. Moreover, on the foundation to provide a legal understanding of the harm caused by hateful messages, communication theories [18] and social psychology [19] are also employed. Clearly, the hate speech problem goes way back in time, but there are still issues requiring careful attention and treatment, especially in today's guzzling world of social media and digital content, with the vast and uncontrollable way of information publishing/propagations and the associated audience reactions.

1.1. Related Work: Hate Speech in Social Media and Proposed Algorithmic Solutions

Hate speech has been a pivotal concept both in public debate and in academia for a long time. However, the proliferation of online journalism along with the diffusion of user-generated content and the possibility of anonymity that it allows [20,21] has led to the increasing presence of hate speech in mainstream media and social networks [22,23].

During the recent decades, media production has often been analyzed through the lens of citizen participation. The idea of users' active engagement in the context of main-

stream media was initially accompanied by promises of enhancing democratization and strengthening bonds with the community [24,25]. However, the empirical reality of user participation was different from the expectations, as there is lots of dark participation, with examples ranging from misinformation and hate campaigns to individual trolling and cyberbullying; a large variety of participation behaviors are evil, malevolent, and destructive [22]. Journalists identify hate speech as a very frequently occurring problem in participatory spaces [8]. Especially comments, which are considered an integral part of almost every news item [26], have become an important section for hate speech spreading [27].

Furthermore, an overwhelming majority of journalists argue that they frequently come upon hate speech towards journalists in general, while most of them report a strong increase in hate speech personally directed at them [28]. When directed at professionals, hate speech can cause negative effects both on journalists themselves and journalistic work: it might impede their ability to fulfill their duties as it can put them under stark emotional pressure, trigger conflict into newsrooms when opinions diverge on how to deal with hateful attacks or even negatively affect journalists' perception of their audience [28]. Hence, not rarely, professionals see users' contributions as a necessary evil [27] and are compelled to handle a vast amount of amateur content in tandem with their other daily tasks [29].

To avoid problems, such as hate speech, and protect the quality of their online outlets, media organizations adopt policies that establish standards of conduct and restrict certain behaviors and expressions by users [30]. Community managers are thus in charge of moderating users' contributions [31], by employing various strategies for supervising, controlling, and enabling content submission [32]. When pre-moderation is followed, every submission is checked before publication and high security is achieved. However, this method requires considerable human, financial, and time resources [27]. On the other hand, post-moderation policies lead to a simpler and more open approach but can lower the quality [33], exposing the platform to ethical and legal risks. Apart from manual moderation, some websites utilize artificial intelligence techniques to tackle this massive work automatically [34], while others implement semi-automatic approaches that assist humans through the integration of machine learning into the manual process [35].

The automation of the process of hate speech detection relies on the training and evaluation of models, using annotated corpora. The main approaches include lexicon-based term detection and supervised machine learning. Lexicons contain a list of terms, along with their evaluation concerning the relation to hate speech. The terms are carefully selected and evaluated by experts on the field, and they need to be combined with rule-based algorithms [36–38]. Such algorithms are based on language-specific syntax and rules. Computational models such as unsupervised topic modeling can lead to insight regarding the most frequent terms that allow further categorization of the hate-related topics [39,40]. In supervised machine learning approaches, models are trained using annotated corpora. Baseline approaches rely on bag-of-words representations combined with machine learning algorithms [36–41]. More recent methods rely on deep learning and word embeddings [42,43]. The robustness of a supervised machine learning algorithm and its ability to generalize for the detection of hate in unseen data relies on the retrieval of vast amounts of textual data.

Big data analytics of social media contents is an emerging field for the management of the huge volumes that are created and expanded daily [44]. Most social media services offer dedicated application programming interfaces (APIs) for the collection of posts and comments, to facilitate the work of academics and stakeholders. Using a dedicated API, or a custom-made internet scraper makes it easy to retrieve thousands of records automatically. Twitter is the most common choice, due to the ease-of-use of its API, and its data structure that makes it easy to retrieve content relevant to a specific topic [36,37,41]. While textual analysis is the core of hate speech detection, metadata containing information about the record (e.g., time, location, author, etc.) may also contribute to model performance.

Hate speech detection cannot be language-agnostic, which means that a separate corpus or lexicon and methodology needs to be formed for every different language [36,37,45]. Moreover, a manual annotation process is necessary, which, inevitably introduces a lot of human effort, as well as subjectivity [36]. Several annotation schemes can be found in literature, differing in language, sources (e.g., Twitter, Facebook, etc.), available classes (e.g., hate speech, abusive language, etc.), and ranking of the degree of hate (e.g., valence, intensity, numerical ranking, etc.) [37]. The selected source itself may influence the robustness of the algorithmic process. For instance, Twitter provides a maximum message length, which can affect the model fitting in a supervised training process [36]. Multi-source approaches indicate the combination of different sources for analytics [46]. In [47] for example, Twitter data from Italy are analyzed using computational linguistics and the results are visualized through a Web platform to make them accessible to the public.

1.2. Project Motivation and Research Objectives

Based on the preceding analysis, there is missing a multilingual hate-speech detection (and prevention) web-service, which individuals can utilize for monitoring informatory streams with questionable content, including their own user-generated content (UGC) posts and comments. More specifically, the envisioned web environment targets to offer an all-in-one service for hate speech detection in text data deriving from social channels, as part of the Preventing Hate against Refugees and Migrants (PHARM) project. The main goal of the PHARM project is to monitor and model hate speech against refugees and migrants in Greece, Italy, and Spain to predict and combat hate crime and also counter its effects using cutting-edge algorithms. This task is supported via intelligent natural language processing mechanisms that identify the textual hate and sentiment load, along with related metadata, such as user location, web identity, etc. Furthermore, a structured database is initially formed and dynamically evolving to enhance precision in subsequent searching, concluding in the formulation of a broadened multilingual hate-speech repository, serving casual, professional, and academic purposes. In this context, the whole endeavor should be put into test through a series of analysis and assessment outcomes (low-/high-fidelity prototypes, alpha/beta testing, etc.) to monitor and stress the effectiveness of the offered functionalities and end-user interface usability in relation to various factors, such as users' knowledge and experience background. Thus, standard application development procedures are followed through the processes of rapid prototyping and the anthropocentric design, i.e., the so-called logical-user-centered-interactive design (LUCID) [48–52]. Therefore, audience engagement is crucial, not only for communicating and listing the needs and preferences of the targeted users but also for serving the data crowdsourcing and annotating tasks. In this perspective, focusing groups with multidisciplinary experts of various kinds are assembled as part of the design process and the pursued formative evaluation [50–52], including journalists, media professionals, communication specialists, subject-matter experts, programmers/software engineers, graphic designers, students, plenary individuals, etc. Furthermore, online surveys are deployed to capture public interest and people's willingness to embrace and employ future Internet tools. Overall, following the above assessment and reinforcement procedures, the initial hypothesis of this research is that it is both feasible and innovative to launch semantic web services for detecting/analyzing hate speech and emotions spread through the Internet and social media and that there is an audience willing to use the application and contribute. The interface can be designed as intuitively as possible to achieve high efficiency and usability standards so that it could be addressed to broader audiences with minimum digital literacy requirements. In this context, the risen research questions (RQ) elaborated to the hypotheses are as follows:

RQ1: Is the PHARM interface easy enough for the targeted users to comprehend and utilize? How transparent the offered functionalities are?
RQ2: What is the estimated impact of the proposed framework on the journalism profession and the anticipated Web 3.0 services? Are the assessment remarks related to the Journalism profession?

2. Materials and Methods

As a core objective of the PHARM project is to build a software environment for querying, analyzing, and storing multi-source news and social media content focusing on hate speech against migrants and refugees, a set of scripts for Natural Language Processing (NLP) has been developed, along with a web service that enables friendly user interaction. Let the former be called the PHARM Scripts, the latter the PHARM Interface, and both of them the PHARM software. All these implementations are constantly elaborated and updated as the project evolves, the source code of the PHARM Scripts, along with the required documentation, is publicly available as a GitHub repository (http://github.com/thepharmproject/set_of_scripts, accessed on 18 March 2021), while the PHARM Interface has the form of a website (http://pharm-interface.usal.es, accessed on 18 March 2021). The detailed documentation of the algorithms is out of the scope of the current work, so only a brief presentation of the relevant functionality follows. Comprehensive documentation and use instructions for the interface are available online (http://pharm-interface.usal.es/instructions, accessed on 18 March 2021) in English, Greek, Italian and Spanish.

2.1. Data Collection

The core outcome of the PHARM software concern a multi-source platform for the analysis of unstructured news and social media messages. On the one hand, it is very important for hate speech texts to include both data (texts of the news or social media messages) and metadata (location, language, date, etc.). On the other hand, the diversity of the sources is unquestionable and thus, mandatory. Therefore, several sources have been selected for the collection of content related to hate speech, while all the required technical implementations have been made to collect the necessary data from these sources. The sources include websites in Greek, Italian, Spanish as well as Twitter, YouTube, and Facebook and concern articles, comments, tweets, posts, and replies. The list of the sources was initialized and updated by the media experts. The list of the sources includes 22 Spanish, 12 Italian, and 16 Greek websites that are prone to publishing hate speech content in the articles or the comments section. Site-specific scripts for scraping have been developed for the collection of semi-structured content (including the accompanying metadata) from the proposed websites, while content from open Facebook groups and pages, as well as websites that are not included in the list, are supported using a site-agnostic scraping method. Tweets are gathered using a list of hashtags and filters containing terms relevant to anti-immigration rhetoric and YouTube comments are collected using search queries relevant to immigration. To store, query, analyze and share news and social media messages, PHARM software adopts a semi-structured format based on JSON (JavaScript object notation), adapted to media features.

2.2. Data Format

Taking into account the requirements of the project (i.e., the use of some relevant extra information for hate speech analysis), the sources that are used for scraping content (i.e., website articles and comments, YouTube comments, tweets), interoperability and compatibility considerations for importing and exporting data between the PHARM Interface, PHARM Scripts, and third-party applications, some general specifications for the data format have been set. The main field is the text (i.e., content), accompanied by the id, annotations, and meta fields. The meta field is a container that includes all metadata. A minimum set of metadata is used for all platforms (i.e., type, plang, pdate, phate, psent, pterms, ploc). These fields are found for all records across different sources. Table 1 presents the proposed data scheme.

Table 1. The common fields of the specified data format.

Field	Description
id	unique identifier
text	content
annotations	hate speech and sentiment annotations
meta/type	type of text (tweet, article, post, comment, etc.)
meta/plang	language detection via PHARM Scripts
meta/pdate	datetime estimation via PHARM Scripts
meta/phate	hate speech detection via PHARM Scripts
meta/psent	sentiment analysis via PHARM Scripts
meta/pterms	frequent terms collection via PHARM Scripts
meta/ploc	geolocation estimation via PHARM Scripts
meta/meta	unsorted metadata

In the cases of web scraping, metadata depends on the available data provided by each site, whereas for YouTube comments and tweets, where the corresponding API is used, specific metadata have been selected and are stored along with the text. Table 2 demonstrates the fields that are used for data that originate from the Twitter and YouTube social media platforms.

Table 2. The metadata fields that are exploited for the YouTube and Twitter records.

Twitter	YouTube
tweet id	comment id
is retweet	reply count
is quote	like count
user id	video id
username	video title
screenname	channel
location	video description
follower count	author id
friend count	author name
date	date

2.3. Data Analysis

The most notable analysis methods that are used in the PHARM software concern date, time, and geolocation estimation, language detection, hate speech detection, and sentiment analysis. Various software libraries have been deployed for implementing the supported analysis methods, along with custom algorithms that have been developed specifically for the PHARM software. A brief description of these methods follows.

Language Detection: The PHARM software mainly processes text produced in Greek, Italian, and Spanish languages but many of the sources may contain texts in other languages or local dialects [53]. To work with these three national languages, a procedure to detect the language of the media text when it is not properly declared has been specified. An ensemble approach for improved robustness is adopted, querying various language detection libraries simultaneously. Amongst the used libraries are the textblob and googletrans Python libraries [54,55].

Geolocation Estimation: Geolocation identification of the collected texts is considered useful for analysis [53]. Therefore, a method for detecting geolocation from text data has been implemented. Named entities are extracted from texts and geocoded i.e., the geographical coordinates are retrieved for the found entities. The named entities include geopolitical entities (GPE) (i.e., countries, cities, states), locations (LOC) (i.e., mountains, bodies of water), faculties (FAC) (buildings, airports, highways, etc.), organizations (ORG) (companies, agencies, institutions, etc.). For this method, the Nominatim geocoder, along with openstreetmap data are used [56].

Datetime Estimation: Besides location and language, when metadata is available, relevant extra information for hate speech analysis can be used. Some of this extra information, such as date or time, may be available in different formats, introducing the necessity of standardization. Therefore, a method for detecting and standardizing date and time information from meta- and text- data has been implemented. A couple of Python libraries (e.g., dateparser, datefinder, and parsedatetime) are exploited for detecting datetime objects in texts. This is based on metadata analysis, where date information is commonly present. If datetime detection fails for the metadata, the same workflow is applied to the text data.

Hate Speech Detection: Undoubtedly, hate speech detection is a core algorithmic asset for the project. Therefore, a couple of methods for detecting hate speech have been implemented, based on both an unsupervised and a supervised approach. The former concerns a lexicon-based method relying on a dictionary containing static phrases, along with dynamic term combinations (i.e., adjectives with nouns), while the latter refers to a machine learning procedure. For both methods, a language model is loaded (according to the language of the text) and common normalization practices are taking place (lowercasing, lemmatization, stop-word and punctuation removal). In the first case, the targeted terms are being searched and the text, while in the second, a recurrent neural network (RNN) undertakes the task of detecting hate speech [41,57]. For this reason, a pretrained tokenizer and a deep network consisting of an embedding layer, a gated recurrent unit (GRU) layer, and a fully connected layer are deployed. The models and the tokenizer have been trained using the Keras framework [58].

Sentiment Analysis: Similarly, two methods for sentiment analysis in the context of hate speech against refugees have been embedded in the interface [45,59]. These follow the same concepts as in hate speech detection but exploiting different lexicons and training data. The unsupervised method adopts many key aspects of the SentiStrength algorithm, such as the detection of booster, question, and negating words [60]. The supervised model for sentiment analysis follows the same architecture as the one for hate speech detection, trained on a different corpus.

Topic Modeling: The lexicons for both hate speech and sentiment analysis were developed by a team of experts in the field of journalism, communication, and media. To facilitate the process with automated text analysis, exploratory content processing techniques for topic modeling and entity collection have been deployed as well. The dictionaries have been built using frequent words, boosted by entity extraction based on ter frequency (TF) for absolute entity counting and term frequency-inverse document frequency (TF-IDF) for proportional counting, showing how important an entity is for the document or even the entire corpus [39,59].

2.4. Project Analysis and Usability Evaluation

The defined research questions and the elongated conclusions were supported by an empirical survey regarding the evaluation of the developed interface, while its statistical results are presented in the respective section of the manuscript. In this context, a multifactor questionnaire was formulated, targeting to record the users' opinions about the web interface and the corresponding functionalities. It has to be noted that in this section, an overview of the questionnaire is exhibited, along with the survey identity, while a detailed description can be accessed in the manuscript appendix. The evaluation process was categorized into 8 major factors, namely the efficiency (7 items), usability (12 items), learnability (5 items), satisfaction (11 items), navigation (4 items), content (6 items), interactivity (5 items) and design (3 items) of the web interface. Furthermore, the participants were called to answer about the area that the web interface usage could cover according to their interests (5 items) and the scope of utilization in public information and awareness (6 items). Taking into account the main functionalities of the web interface, a stand-alone question was posed regarding the assessment of the contribution of the project towards the identification of hate speech mechanisms and sentiment loads in text data. All the aforementioned metrics were measured on a 5-level Likert scale, ranging from 1—strongly

disagree to 5—strongly agree. The demographic questions that were involved in the survey addressed gender (male, female, other, no answer), age (18–22, 23–30, 31–40, 41–50, over 50), education (high school, vocational learning, bachelor, master, Ph.D.), computer familiarity (Likert scale 1–5), Internet familiarity (Likert scale 1–5), news awareness (Likert scale 1–5) and the profession (in 9 categories) of the participants. However, a dedicated binary inquiry was inserted recording if the participant works/ has worked as a journalist (yes-no), because of the additive value in the assessment of the web service that focuses on hate speech detection. Table 3 summarizes the set of questions that were implicated in the current survey.

Table 3. Overview of the formulated questionnaire.

#	Questions/Factors	Measure
1	Efficiency-7 items	Likert Scale 1–5
2	Usability-12 items	Likert Scale 1–5
3	Learnability 5 items	Likert Scale 1–5
4	Satisfaction-11 items	Likert Scale 1–5
5	Navigation-4 items	Likert Scale 1–5
6	Content-6 items	Likert Scale 1–5
7	Interactivity-5 items	Likert Scale 1–5
8	Design-3 items	Likert Scale 1–5
9	"Use for" scenarios-5 items	Likert Scale 1–5
10	Public Information and Awareness-6 items	Likert Scale 1–5
11	Contribution to hate speech/ sentiment detection	Likert Scale 1–5
12	Gender	Male/Female/Other/No answer
13	Age	18–22, 23–30, 31–40, 41–50, 50+
14	Education	Highschool, Vocational Learning, Bachelor, Master, PhD
15	Computer Familiarity	Likert Scale 1–5
16	Internet Familiarity	Likert Scale 1–5
17	News Awareness	Likert Scale 1–5
18	Profession	9 Categories
19	Working/has worked as Journalist	Binary Yes-No

The survey was conducted mainly via the social media channels of the authors, while the final number of gathered responses reached $n = 64$. The temporal length upon the completion of the survey ranged from 7 to 10 min, while in this duration the participants had to navigate and interact with the website interface and at the same time to answer the projected inquiries regarding its evaluation. The moderate number of responses can be justified on the multidisciplinary nature of the conducted research, which prerequisites more expertise and of course more time since several tasks were involved during the assessment process. Nevertheless, the aforementioned argument favors the reliability of the evaluation results because of the volunteered engagement of the 64 users in the survey. Finally, it has to be highlighted that this is the first time that the project is assessed, aiming at preliminary evaluation remarks for further optimizing crucial aspects of the interface based on the participants' opinions. A reliability test was conducted on the questionnaire, based on Cronbach's alpha, revealing the respective coefficient a = 0.87, therefore supporting confident statistical results in the following section. Table 4 presents the basic demographic information of the respondents.

Table 4. Demographic Information of Participants.

#	Question	Answers-Distribution
1	Gender	Male (28.1%), Female (71.9%)
2	Age	18–22 (12.5%), 23–30 (43.8%), 31–40 (34.4%), 41–50 (4.7%), 50+ (4.7%)
3	Education	Highschool (14.1%), Vocational Learning (1.6%), Bachelor (31.3%), Master (45.3%), PhD (7.8%)
4	Computer Familiarity	1 (3.1%), 2 (29.7%), 3 (20.3%), 4 (34.4%), 5 (12.5%)
5	Internet Familiarity	1 (1.6%), 2 (18.8%), 3 (9.3%), 4 (54.7%), 5 (15.6%)
6	News Awareness	1 (0%), 2 (3.1%), 3 (21.9%), 4 (37.5%), 5 (37.5%)
7	Working/has worked as Journalist	Yes (46.9%), No (53.1%)

During the survey preparation, all ethical approval procedures and rules suggested by the "Committee on Research Ethics and Conduct" of the Aristotle University of Thessaloniki were followed.

3. Results

The results of the current research concern the presentation of the functionality and usability, along with the multi-faceted evaluation of the implemented web interface.

3.1. The Implemented Web Interface

The PHARM Interface serves as the front-end of the PHARM software. It is the graphical interface for exposing data and functionality to the users and relies on the back-end, which consists of the PHARM scripts. For the development of the interface, the Python web framework Flask has been used. The choice is justified, as the NLP and data analysis scripts are also written in Python and, following this approach, all the functionality of the interface can be included within a common software project. The graphical user interface (GUI) has been mainly designed in Bootstrap, a popular HTML, CSS, and JavaScript library. Additional HTML, CSS, and JavaScript blocks have been added where needed. The Flask project has been deployed on a virtual machine and is served using the Waitress, a production-quality pure-Python web server gateway interface (WSGI) with very acceptable performance. The PHARM Interface is accessible at http://pharm-interface.usal.es (accessed on 18 March 2021). The home screen of the Interface gives some basic information about the PHARM project and provides a starting point for accessing the supported NLP methods (Figure 1).

Let us further describe the software by analyzing the types of users and the actions that have been formally specified. The core functions of the Interface are five: search, analyze, scrape, annotate, and submit, whereas two types of users have been defined: the visitor and the contributor. A visitor can search and analyze hate speech data, while the contributor can also scrape, annotate and submit relevant content. The functions that are available to all users can be considered as public, while the rest as private. The private functionality is only accessible by registered users which are intended to be media professionals. Figure 2 demonstrates the succession of all functions in a single workflow, divided into the two aforementioned groups. As the current work focuses on the public functionality of the interface, only a brief description of the private functions is given.

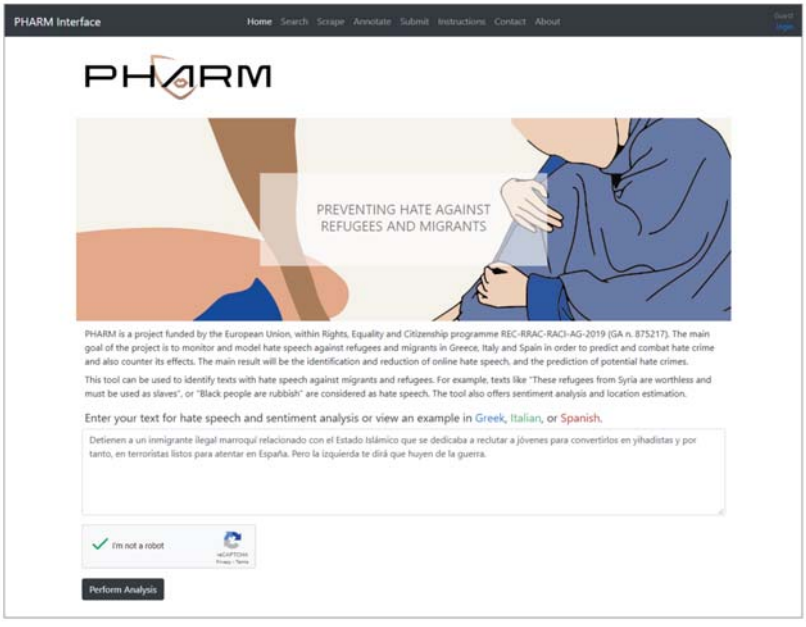


Figure 1. The home screen of the Preventing Hate against Refugees and Migrants (PHARM) web interface.

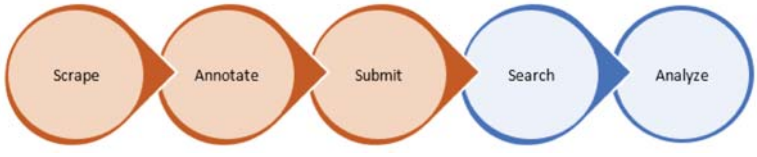


Figure 2. The private (orange) and public (blue) functionality of the PHARM Interface.

Search: One of the main functionalities of the interface is the navigation through the hate speech records contained in the database. The user can view records for any supported language (English, Greek, Italian, Spanish) or, additionally, filter the results by applying a variety of filters. The available filters are:

- Source selection (tweets, Facebook posts and replies, website articles and comments).
- Date and time selection (show results inside a specific period).
- Annotation filtering (hate/no hate, positive/neutral/negative sentiment).
- Keyword filtering (a search query for finding occurrences to texts).

The user can preview the records as a list, download them as a CSV or a JSON file, or display detailed information for each item. The search results can be viewed and downloaded either in the "simple" or the "scientific" form, disabling or enabling the presence of metadata, respectively. Figure 3 presents the search results screen of the interface.

Analyze: When a record is selected, or a text is placed on the home screen, a detailed report appears. The location is marked on a map and the results of various text analysis algorithms are presented with graphics (icons, bars, etc.). The results concern hate speech detection and sentiment analysis (for both unsupervised and supervised classification methods), frequent entity detection, and geolocation estimation. Figure 4 depicts the analysis screen of the PHARM interface.

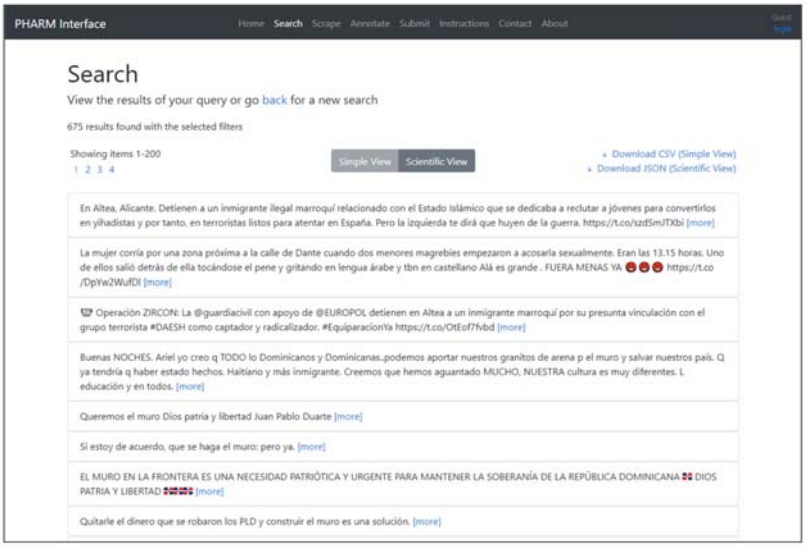


Figure 3. The search results screen of the PHARM Interface.

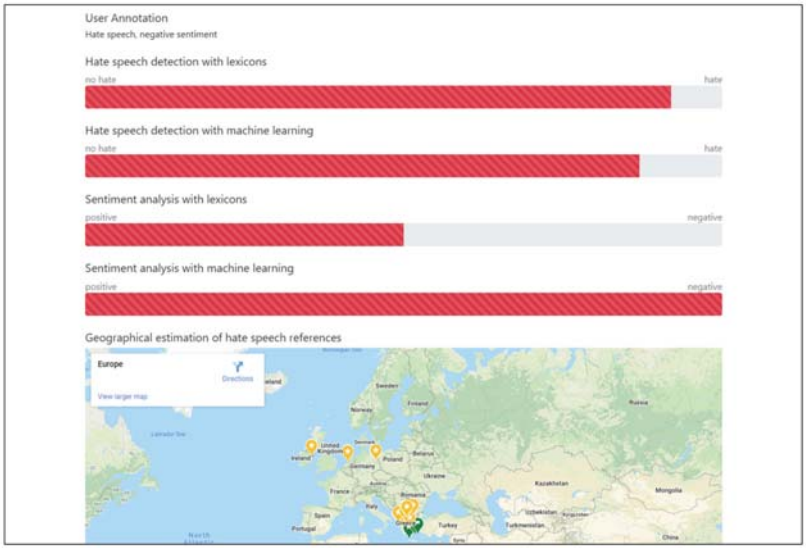


Figure 4. The analysis screen of the PHARM Interface.

Scrape: The PHARM interface enables the mass-collection of text data from two popular platforms: Twitter and YouTube. A user can collect hate speech data from Twitter, by selecting language (Greek, English, Italian, or Spanish) and invoking the process. The tweets are collected based on language-specific lexicons that have been developed in the context of the project. The process stops after a user-configurable time interval and a link is provided for downloading a JSON file that contains the data. These data may be further used for annotation, submission to the PHARM database, or any other NLP task. In the case of YouTube, instead of selecting the language, a search query should be set. The search

query can include individual search terms or a combination of them, separated by a comma. The resulting data can be downloaded as a CSV or JSON file.

Annotate: The annotation process is powered by the Doccano tool [61]. Doccano is an annotation management system for text data and can be used for developing datasets for facilitating classification, entity tagging, or translation tasks. In the context of the PHARM project, it is used for text classification and each record should be labeled with specific tags denoting hate speech and sentiment load.

Submit: Data entry can be executed either one by one or massively. Concerning the first method, the user should set all data (text) and metadata (source, language, date, hate, sentiment, etc.) via the corresponding input forms (i.e., text fields, radio buttons, etc.). If data are already formed appropriately, they can be imported as a JSON file too.

3.2. Analysis and Usability Evaluation Results

The web interface was developed during the last year, while critical alpha evaluation tests were conducted inside the research team. Furthermore, the implemented interface was subjected to a beta assessment process by experts in the scientific fields of web design, web graphics, etc., aiming at the detection and correction of problematic aspects/ flaws at an early stage. Consequently, the final step of the evaluation was the conducted broadened empirical survey via the formulated questionnaire of Section 2.4, while in this section the extracted results are presented. Specifically, Figure 5 exhibits the responses regarding the suitability of the web interface towards the hate speech and sentiment loads detection mechanisms in text data, gathering 75% of agree/strongly agree evaluation score.

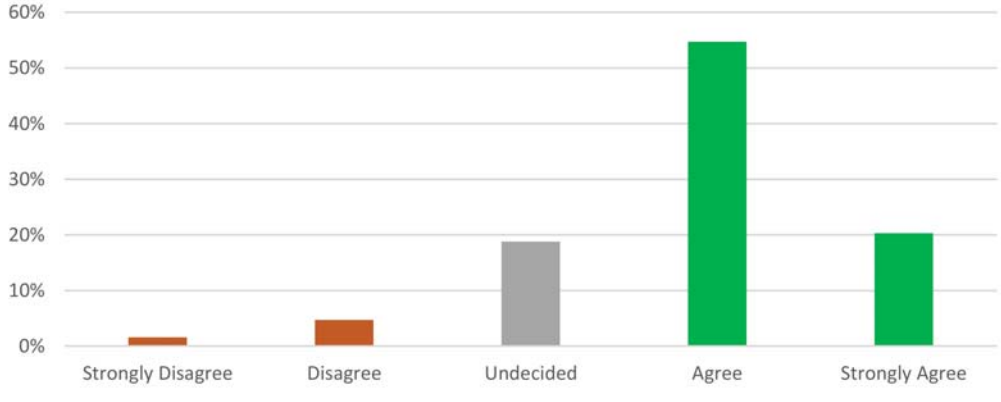


Figure 5. Responses to the suitability of the PHARM interface.

As Table 5 presents, the eight evaluation factors are constituted by various numbers of items, therefore for homogeneity and statistical purposes, the average scores of the Likert-scaled items were computed for each metric. Table 5 exhibits the mean and standard deviation values of the assessment factors based on the responses of the $n = 64$ participants. The results showed that in all aspects the web interface grades are prone to four with a lower to one standard deviation, therefore indicating a well-designed web interface, with increased usability, presenting comprehensive content, navigation, and interaction mechanisms, and being easy-to-learn. Of course, these general results are further processed towards the determination of inner class correlations and group scores differentiations to address the defined research questions of the survey.

Table 5. Statistical values of evaluation factors.

#	Factor	Mean	Standard Deviation
1	Efficiency	3.72	0.73
2	Usability	3.95	0.64
3	Learnability	3.97	0.68
4	Satisfaction	3.71	0.73
5	Navigation	3.93	0.82
6	Content	3.74	0.55
7	Interactivity	3.85	0.67
8	Design	3.66	0.88

One of the main concerns for the effective engagement of users into a web interface is their knowledge background, possible previous experience in similar services, etc. Therefore, while addressing RQ1, the evaluation metrics were examined in relation with computer and Internet familiarity of the implicated users, towards the extraction of meaningful results. Taking into consideration that the factors of computer familiarity and Internet familiarity are answered in a 5-level Likert scale, the subjective judgment of knowledge background is unavoidably inserted in the statistical analysis. Because of possible error propagation due to the moderate number of participants and also the preliminary nature of the conducted survey, the responses in these two factors are grouped into two major categories. Specifically, the recoded categories are poor familiarity (including Answers 1 and 2) and good familiarity (including Answers 4 and 5), leaving out the moderate/ ambiguous level of computer and Internet familiarity (Level 3 in Likert scale), therefore functioning in a binary mode. Taking into consideration the continuous-form variables of the evaluation factors and the nominal two-scaled grouping of computer and Internet familiarity items, the statistical analysis proceeded into independent samples *t*-test methods, in order to compute the average scores differentiations of the assessment values into the formulated groups of participants.

Figure 6 graphically exhibits the crosstabulation matrices of average evaluation scores of the groups, while Table 6 presents the calculated values of the conducted *t*-tests, with significance level a = 0.05. In this context, statistically significant differentiations in average scores between groups were computed for usability, learnability, navigation in the computer familiarity groups, while only for the first two ones in the Internet familiarity groups.

Because of the specific contribution of the PHARM Project in hate speech detection in textual data, affecting public news and awareness, the second research question (RQ2) refers to the assessment of the web interface from participants that work (or have worked) in journalism compared to simple users with other professions. For this reason, a dedicated independent samples *t*-test was conducted for the evaluation scores of these two groups (answering YES if they work/have worked as journalists and NO if not the case).

Figure 7 presents the average values of the eight evaluation factors for the two subsets of participants, while Table 7 exhibits the related *t*-test outputs, again in a = 0.05 significance level. As can be observed, there was statistical significant difference in the average scores of the two groups only for the efficiency evaluation factor.

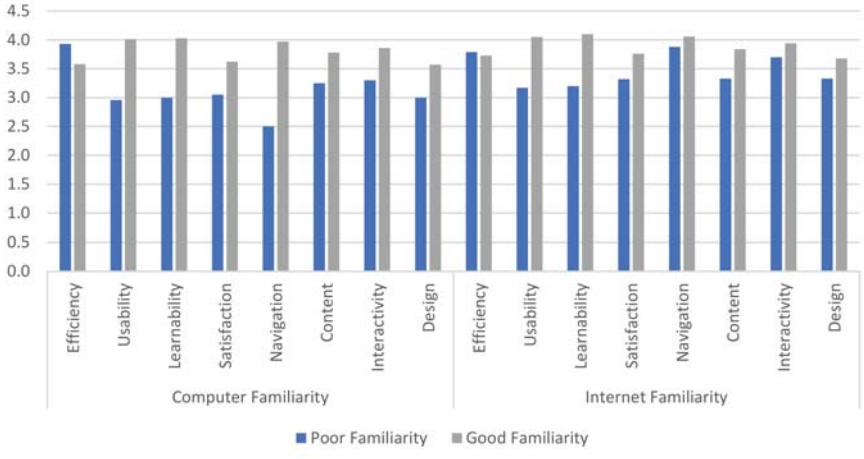


Figure 6. Average evaluation scores of the two groups for computer and internet familiarity variables.

Table 6. T-tests results for correlation between the evaluation factors and the groups of computer and Internet familiarity variables.

#	Factor	Computer Familiarity		Internet Familiarity	
		t-Value	*p*-Value	t-Value	*p*-Value
1	Efficiency	0.598	0.554	0.105	0.917
2	Usability	−2.514	0.018 *	−2.067	0.045 *
3	Learnability	−2.283	0.030 *	−2.217	0.032 *
4	Satisfaction	−1.015	0.318	−0.821	0.416
5	Navigation	−2.265	0.031 *	−0.349	0.728
6	Content	−1.267	0.215	−1.347	0.185
7	Interactivity	−1.231	0.228	−0.525	0.602
8	Design	−0.764	0.451	−0.517	0.608

* Statistically significant difference between groups at a = 0.05 significance level

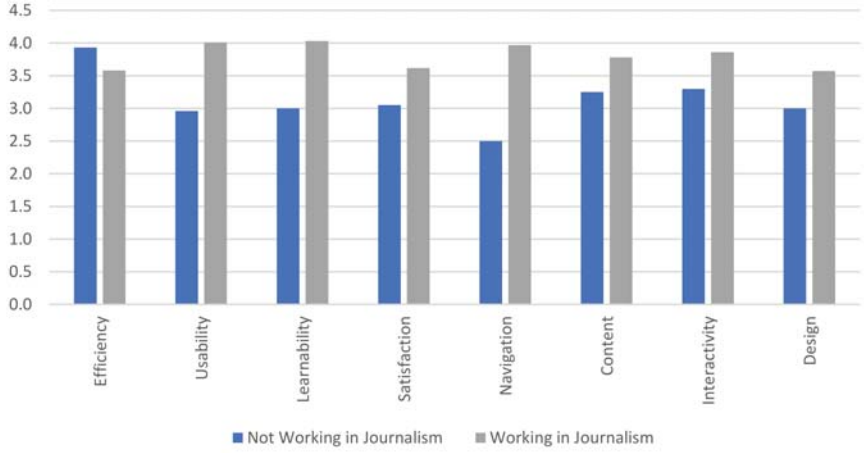


Figure 7. Average evaluation scores of groups in Journalism variable.

Table 7. Independent samples *t*-test of evaluation scores for the groups of working/ have worked as journalists or not.

#	Factor	t-Value	*p*-Value
1	Efficiency	2.230	0.029 *
2	Usability	−0.399	0.691
3	Learnability	0.096	0.924
4	Satisfaction	0.275	0.784
5	Navigation	−0.033	0.974
6	Content	0.425	0.672
7	Interactivity	0.750	0.456
8	Design	−0.278	0.782

* Statistically significant difference between groups at a = 0.05 significance level

4. Discussion

Overall, as Figure 6 presents, the PHARM interface was positively evaluated by the engaged participants (75%), while special attention was given to the remaining 25% concerning possible problems or negative aspects and functionalities of the platform, for subsequent developed versions. While referring to RQ1, the second column of Table 6 indicates that there is statistical significant difference between the mean scores of the groups of computer familiarity with respect to usability ($p = 0.018$), learnability ($p = 0.030$), and navigation ($p = 0.031$) evaluation factors of the interface. This fact is also validated in Figure 6, since the average values for the computer poor familiarity group are substantial lower versus the good familiarity one, in the factor of usability (2.96 vs. 4.01), learnability (3.00 vs. 4.03) and navigation (2.50 vs. 3.97). These results imply that amateurs in computer science participants confronted potential difficulties while navigating or learning how to utilize the web interface. In this context, some modifications are necessary for further optimizing the end-user interface to become more comprehensive, with increased usability. Nevertheless, for the rest of five evaluation factors, there was no statistically significant difference between the groups of computer familiarity, which is in accordance with the similar average values in these cases of Figure 6. Consequently, the conducted tests indicated towards the users' satisfaction and web service efficiency, along with the inclusion of adequate content, design, and interactivity mechanisms.

Almost the same results were retrieved for the two groups of Internet familiarity (RQ2), since there was statistically significant difference only for usability ($p = 0.045$) and learnability ($p = 0.032$) metrics (without any difference in navigation). The average evaluation scores of Internet poor familiarity group were substantially lower compared to the good familiarity one for usability (3.17 vs. 4.05) and learnability (3.20 vs. 4.09), while there were no crucial differentiations for the remaining six evaluation factors. Taking into consideration the aforementioned results, the web interface was, in general, positively evaluated by all participants for most of its aspects, while specific actions are required in further optimizations/evolvement of the platform, to address the low usability and learnability scores for the less technologically experienced users. For instance, short "how-to" videos and simplified versions of manuals are already discussed among the research team members, to address the usability, navigation, and learnability deficiencies for potential amateur users.

With regard to RQ2, the extracted *p* values of Table 7 indicate that there is a statistically significant difference of evaluation scores between the two groups only for the factor of the efficiency ($p = 0.029$) of the web platform, while the exact average values are 3.51 for the subset who work/ have worked in journalism compared to 3.91 for those who have no relationship with it. This fact implies that the first group remains somehow skeptical about the effectiveness of the web service towards the detection of hate speech and emotional load in text, which mainly relies on human-centric approaches due to the implicated subjectivity. Therefore, the integrated natural language processing modules will be further evolved to achieve maximum precision, persuading for applicability of the innovative automations without human intervention. However, it has to be highlighted that in all other assessment

metrics there was no substantial difference in the average evaluation scores, validating the high-quality content, design, navigation mechanisms, etc., either for professionals or simple users (with scores usually close to four for both groups).

The web service that has been presented in the current paper is part of a software framework for the collection and analysis of texts from several social media and websites, containing hate speech against refugees. The web service covers the functionality of web scraping, annotating, submitting annotated content and querying the database. It supports multi-language and multi-source content collection and analysis. This allows the formulation of a comprehensive database that can lead to the development of generalized hate speech and sentiment polarity modeling. This is expected to contribute significantly in the enhancement of semantic aware augmentation of unstructured web content. Future plans include an in-depth evaluation of state-of-the-art technologies in the big data volumes that are collected and annotated constantly through the PHARM software. Providing the functionality and the database online, makes it accessible to the public and allows more people to get involved. The results of the formative evaluation that are presented validate the appeal of the project to the target audience, and provide important feedback for the improvement of future versions. Subsequent larger scale and more generalized evaluation phases will follow, according to the adopted human-centered LUCID design.

Author Contributions: Conceptualization, C.A.-C., A.V., and C.D.; methodology, L.V., N.V., and R.K.; software, L.V., and N.V.; formal analysis, R.K.; investigation, T.S., M.M., and R.K.; resources, L.V., N.V., and R.K.; data curation, T.S., and M.M.; writing—original draft preparation, L.V., N.V., R.K., T.S., M.M., C.D., and A.V.; writing—review and editing, L.V., N.V., R.K., T.S., M.M., C.D., and A.V.; visualization, L.V., N.V., and R.K.; supervision, A.V., C.D. and C.A.-C.; project administration, A.V., C.A.-C., and C.D.; funding acquisition, C.A.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Right Equality and Citizenship Programme (2014–2020). REC-RRAC-RACI-AG-2019 Grant Agreement 875217.

Data Availability Statement: All data that are not subjected to institutional restrictions are available through the links provided within the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Matsiola, M.; Dimoulas, C.A.; Kalliris, G.; Veglis, A.A. Augmenting User Interaction Experience Through Embedded Multimodal Media Agents in Social Networks. In *Information Retrieval and Management*; IGI Global: Hershey, PA, USA, 2018; pp. 1972–1993.
2. Siapera, E.; Veglis, A. *The Handbook of Global Online Journalism*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018.
4. Dimoulas, C.; Veglis, A.; Kalliris, G. Application of mobile cloud based technologies in news reporting: Current trends and future perspectives. In *Joel Rodrigues*; Lin, K., Lloret, J., Eds.; Mobile Networks and Cloud Computing Convergence for Progressive Services and Applications; IGI Global: Hershey, PA, USA, 2014; Chapter 17; pp. 320–343.
5. Dimoulas, C.A.; Symeonidis, A.L. Syncing Shared Multimedia through Audiovisual Bimodal Segmentation. *IEEE MultiMedia* **2015**, *22*, 26–42. [CrossRef]
6. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How in Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173. [CrossRef]
7. Dimoulas, C.A.; Veglis, A.A.; Kalliris, G.; Khosrow-Pour, D.M. Semantically Enhanced Authoring of Shared Media. In *Encyclopedia of Information Science and Technology, Fourth Edition*; IGI Global: Hershey, PA, USA, 2018; pp. 6476–6487.
8. Saridou, T.; Veglis, A.; Tsipas, N.; Panagiotidis, K. Towards a semantic-oriented model of participatory journalism management. Available online: https://coming.gr/wp-content/uploads/2020/02/2_2019_JEICOM_SPissue_Saridou_pp.-27-37.pdf (accessed on 18 March 2021).
9. Cammaerts, B. Radical pluralism and free speech in online public spaces. *Int. J. Cult. Stud.* **2009**, *12*, 555–575. [CrossRef]
10. Fortuna, P.; Nunes, S. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.* **2018**, *51*, 1–30. [CrossRef]
11. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Palo Alto, CA, USA, 25–28 June 2017.
12. Ekman, M. Anti-immigration and racist discourse in social media. *Eur. J. Commun.* **2019**, *34*, 606–618. [CrossRef]

13. Burnap, P.; Williams, M.L. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In Proceedings of the 2014 Internet, Policy & Politics Conferences, Oxford, UK, 15–26 September 2014.
14. Pohjonen, M.; Udupa, S. Extreme speech online: An anthropological critique of hate speech debates. *Int. J. Commun.* **2017**, *11*, 1173–1191.
15. Ben-David, A.; Fernández, A.M. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *Int. J. Commun.* **2016**, *10*, 1167–1193.
16. Olteanu, A.; Castillo, C.; Boy, J.; Varshney, K. The effect of extremist violence on hateful speech online. In Proceedings of the twelfth International AAAI Conference on Web and Social Media, Stanford, CA, USA, 25–28 June 2018.
17. Paz, M.A.; Montero-Díaz, J.; Moreno-Delgado, A. Hate Speech: A Systematized Review. *SAGE Open* **2020**, *10*. [CrossRef]
18. Calvert, C. Hate Speech and Its Harms: A Communication Theory Perspective. *J. Commun.* **1997**, *47*, 4–19. [CrossRef]
19. Boeckmann, R.J.; Turpin-Petrosino, C. Understanding the harm of hate crime. *J. Soc. Issues* **2002**, *58*, 207–225. [CrossRef]
20. Anderson, P. *What Is Web 2.0? Ideas, Technologies and Implications for Education*; JISC: Bristol, UK, 2007.
21. Kim, Y.; Lowrey, W. Who are Citizen Journalists in the Social Media Environment? *Digit. J.* **2014**, *3*, 298–314. [CrossRef]
22. Quandt, T. Dark Participation. *Media Commun.* **2018**, *6*, 36–48. [CrossRef]
23. Schmidt, A.; Wiegand, M. A Survey on Hate Speech Detection using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 3 April 2017; pp. 1–10.
24. Bruns, A. The active audience: Transforming journalism from gatekeeping to gatewatching. In *Making Online News: The Ethnography of New Media Production*; Paterson, C., Domingo, D., Eds.; Peter Lang: New York, NY, USA, 2008; pp. 171–184.
25. Gillmor, D. *We the media. Grassroots Journalism by the People, for the People*; O'Reilly: Sebastopol, CA, USA, 2004.
26. Hanitzsch, T.; Quandt, T. Online journalism in Germany. In *The Handbook of Global Online Journalism*; Siapera, E., Veglis, A., Eds.; Wiley-Blackwell: West Sussex, UK, 2012; pp. 429–444.
27. Singer, J.B.; Hermida, A.; Domingo, D.; Heinonen, A.; Paulussen, S.; Quandt, T.; Reich, Z.; Vujnovic, M. *Participatory Journalism. Guarding Open Gates at Online Newspapers*; Wiley-Blackwell: Malden, MA, USA, 2018. [CrossRef]
28. Obermaier, M.; Hofbauer, M.; Reinemann, C. Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it. *Stud. Commun. Media* **2018**, *7*, 499–524. [CrossRef]
29. Boberg, S.; Schatto-Eckrodt, T.; Frischlich, L.; Quandt, T. The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum. *Media Commun.* **2018**, *6*, 58–69. [CrossRef]
30. Wolfgang, J.D. Pursuing the Ideal. *Digit. J.* **2015**, *4*, 764–783. [CrossRef]
31. Wintterlin, F.; Schatto-Eckrodt, T.; Frischlich, L.; Boberg, S.; Quandt, T. How to Cope with Dark Participation: Moderation Practices in German Newsrooms. *Digit. J.* **2020**, *8*, 904–924. [CrossRef]
32. Masullo, G.M.; Riedl, M.J.; Huang, Q.E. Engagement Moderation: What Journalists Should Say to Improve Online Discussions. *J. Pract.* **2020**, 1–17. [CrossRef]
33. Hille, S.; Bakker, P. Engaging the social news user: Comments on news sites and Facebook. *J. Pract.* **2014**, *8*, 563–572. [CrossRef]
34. Wang, S. Moderating Uncivil User Comments by Humans or Machines? The Effects of Moderation Agent on Perceptions of Bias and Credibility in News Content. *Digit. J.* **2021**, *9*, 64–83. [CrossRef]
35. Risch, J.; Krestel, R. Delete or not delete? Semi-automatic comment moderation for the newsroom. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, Santa Fe, NM, USA, 25 August 2018; pp. 166–176.
36. MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate speech detection: Challenges and solutions. *PLoS ONE* **2019**, *14*, e0221152. [CrossRef]
37. Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Comput. Sci. Rev.* **2020**, *38*, 100311. [CrossRef]
38. Gitari, N.D.; Zhang, Z.; Damien, H.; Long, J. A Lexicon-based Approach for Hate Speech Detection. *Int. J. Multimed. Ubiquitous Eng.* **2015**, *10*, 215–230. [CrossRef]
39. Arcila-Calderón, C.; de la Vega, G.; Herrero, D.B. Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain. *Soc. Sci.* **2020**, *9*, 188. [CrossRef]
40. Arcila-Calderón, C.; Herrero, D.B.; Frías, M.; Seoanes, F. Refugees Welcome? Online Hate Speech and Sentiments in Twitter 2 in Spain during the reception of the boat Aquarius. *Sustainability* **2021**, *13*, 2728. [CrossRef]
41. Arcila-Calderón, C.; Blanco-Herrero, D.; Apolo, M.B.V. Rechazo y discurso de odio en Twitter: Análisis de contenido de los tuits sobre migrantes y refugiados en español/Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish. *Rev. Española Investig. Sociol.* **2020**, *172*, 21–40. [CrossRef]
42. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on Compiler Construction, Austin, TX, USA, 5–6 February 2017; pp. 759–760.
43. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **2018**, *48*, 4730–4742. [CrossRef]
44. Ghani, N.A.; Hamid, S.; Hashem, I.A.T.; Ahmed, E. Social media big data analytics: A survey. *Comput. Hum. Behav.* **2019**, *101*, 417–428. [CrossRef]
45. Sánchez-Holgado, P.; Arcila-Calderón, C. Supervised Sentiment Analysis of Science Topics: Developing a Training Set of Tweets in Spanish. *J. Infor. Technol. Res.* **2020**, *13*, 80–94. [CrossRef]

46. Korkmaz, G.; Cadena, J.; Kuhlman, C.J.; Marathe, A.; Vullikanti, A.; Ramakrishnan, N. Multi-source models for civil unrest forecasting. *Soc. Netw. Anal. Min.* **2016**, *6*, 1–25. [CrossRef]
47. Capozzi, A.T.; Lai, M.; Basile, V.; Poletto, F.; Sanguinetti, M.; Bosco, C.; Patti, V.; Ruffo, G.; Musto, C.; Polignano, M.; et al. Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project. In Proceedings of the 6th Italian Conference on Computational Linguistics, CLiC-it 2019, Bari, Italy, 13–15 November 2019; Volume 2481, pp. 1–6.
48. Dimoulas, C.A. Multimedia. In *The SAGE International Encyclopedia of Mass Media and Society*; Merskin, D.L., Ed.; SAGE Publications, Inc.: Saunders Oaks, CA, USA, 2019.
49. Dimoulas, C.A. *Multimedia Authoring and Management Technologies: Non-Linear Storytelling in the New Digital Media*; Association of Greek Academic Libraries: Athens, Greece, 2015; Available online: http://hdl.handle.net/11419/4343 (accessed on 18 March 2021). (In Greek)
50. Chatzara, E.; Kotsakis, R.; Tsipas, N.; Vrysis, L.; Dimoulas, C. Machine-Assisted Learning in Highly-Interdisciplinary Media Fields: A Multimedia Guide on Modern Art. *Educ. Sci.* **2019**, *9*, 198. [CrossRef]
51. Psomadaki, O.; Dimoulas, C.; Kalliris, G.; Paschalidis, G. Digital storytelling and audience engagement in cultural heritage management: A collaborative model based on the Digital City of Thessaloniki. *J. Cult. Herit.* **2019**, *36*, 12–22. [CrossRef]
52. Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C.; Veglis, A. MAthE the Game: A Serious Game for Education and Training in News Verification. *Educ. Sci.* **2019**, *9*, 155. [CrossRef]
53. Graham, M.; Hale, S.A.; Gaffney, D. Where in the World Are You? Geolocation and Language Identification in Twitter. *Prof. Geogr.* **2014**, *66*, 568–578. [CrossRef]
54. De Vries, E.; Schoonvelde, M.; Schumacher, G. No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Politi. Anal.* **2018**, *26*, 417–430. [CrossRef]
55. Loria, S. Textblob Documentation. Available online: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf (accessed on 18 March 2021).
56. Clemens, K. Geocoding with openstreetmap data. In Proceedings of the GEOProcessing 2015, Lisbon, Portugal, 22–27 February 2015; p. 10.
57. Arcila-Calderón, C.; Amores, J.; Blanco, D.; Sánchez, M.; Frías, M. Detecting hate speech against migrants and refugees in Twitter using supervised text classification. In Proceedings of the International Communication Association's 71th Annual Conference, Denver, CO, USA, 27–31 May 2021.
58. Chollet, F. Keras: The Python Deep Learning Library. Available online: http://ascl.net/1806.022 (accessed on 18 March 2021).
59. Spiliotopoulou, L.; Damopoulos, D.; Charalabidis, Y.; Maragoudakis, M.; Gritzalis, S. Europe in the shadow of financial crisis: Policy Making via Stance Classification. In Proceedings of the 50th Hawaii International Conference on System Sciences (2017), Hilton Waikoloa Village, HI, USA, 4–7 January 2017.
60. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558. [CrossRef]
61. Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. Doccano: Text Annotation tool for Human. 2018. Available online: https://github.com/doccano/doccano (accessed on 18 March 2021).

future internet

MDPI

Article

An AI-Enabled Stock Prediction Platform Combining News and Social Sensing with Financial Statements

Traianos-Ioannis Theodorou [1,*], Alexandros Zamichos [1], Michalis Skoumperdis [1], Anna Kougioumtzidou [1], Kalliopi Tsolaki [1], Dimitris Papadopoulos [1], Thanasis Patsios [2], George Papanikolaou [2], Athanasios Konstantinidis [3], Anastasios Drosou [1] and Dimitrios Tzovaras [1]

[1] Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thessaloniki, Greece; zamihos@iti.gr (A.Z.); skoumpmi@iti.gr (M.S.); annak@iti.gr (A.K.); ktsolaki@iti.gr (K.T.); dpapadop@iti.gr (D.P.); drosou@iti.gr (A.D.); dimitrios.tzovaras@iti.gr (D.T.)
[2] Media2Day Publishing S.A., 15232 Athens, Greece; patsios@media2day.gr (T.P.); gpap@media2day.gr (G.P.)
[3] Department of Electrical Engineering, Imperial College London, London SW7 2AZ, UK; a.konstantinidis16@imperial.ac.uk
* Correspondence: theodorou@iti.gr

Abstract: In recent years, the area of financial forecasting has attracted high interest due to the emergence of huge data volumes (big data) and the advent of more powerful modeling techniques such as deep learning. To generate the financial forecasts, systems are developed that combine methods from various scientific fields, such as information retrieval, natural language processing and deep learning. In this paper, we present ASPENDYS, a supportive platform for investors that combines various methods from the aforementioned scientific fields aiming to facilitate the management and the decision making of investment actions through personalized recommendations. To accomplish that, the system takes into account both financial data and textual data from news websites and the social networks Twitter and Stocktwits. The financial data are processed using methods of technical analysis and machine learning, while the textual data are analyzed regarding their reliability and then their sentiments towards an investment. As an outcome, investment signals are generated based on the financial data analysis and the sensing of the general sentiment towards a certain investment and are finally recommended to the investors.

Keywords: Web 3.0; machine learning; sentiment analysis; portfolio optimization; portfolio management; media industry; social media; model-based trading

Citation: Theodorou, T.-I.; Zamichos, A.; Skoumperdis, M.; Kougioumtzidou, A.; Tsolaki, K.; Papadopoulos, D.; Patsios, T.; Papanikolaou, G.; Konstantinidis, A.; Drosou, A.; et al. An AI-Enabled Stock Prediction Platform Combining News and Social Sensing with Financial Statements. *Future Internet* **2021**, *13*, 138. https://doi.org/10.3390/fi13060138

Academic Editor: Charalampos Dimoulas

Received: 29 April 2021
Accepted: 18 May 2021
Published: 21 May 2021

1. Introduction

The financial services sector provides services in the field of finance to individuals and corporations. This part of the economy consists of a variety of financial firms including finance companies, banks, investment houses and insurance companies. The industry of financial services is probably the most important sector of the economy, leading the world in terms of earnings and equity market capitalization.

Individuals may access financial markets such as stocks and bonds through investment services. Brokers that are either individuals or online services facilitate the buying and selling of securities. Financial advisors are responsible for the assets management as well as the trades of the portfolio assets. Additionally, quant fund is the latest trend of financial advice and portfolio management, with fully automated algorithmic portfolio optimization and trade executions. Quant funds refer to investment funds whose securities are chosen based on the quantitative analysis of numerical data. The analysis is performed by software programs which utilize advanced mathematical and Artificial Intelligence (AI) models. These models deal with several challenging tasks in the financial services sector, such as the accurate prediction of the stock prices. In contrast with traditional methods, quant funds rely on algorithmic or systematically programmed investment strategies. Thus, they

are not human-oriented and their decisions rely only on data analysis. In this way, they can be very useful tools for supporting investors in their investment's decision making. Regarding the stock prices prediction, many methods have been proposed by researchers since the beginning of the stock market [1–7].

Moreover, during the last years with the advent of big data and machine learning in Web 3.0 technologies, financial forecasting and model-based trading have gained growing interest. Web 3.0 is the third generation of Internet services for websites and applications that focus on using a machine-based understanding of data to provide a data-driven and semantic web. While the traditional methods for financial forecasting were based on the analysis of the stock prices of the past, nowadays, the efforts consider a variety of data. More specifically, the huge amount of available data can be found on social media and from the media industry (articles and posts) that may be related with significant societal and political events. These events, as well as the extracted information from their sentiment analysis, may reflect the stock prices and trends of certain assets. Additionally, investors tend to follow the general sentiment on a specific topic, something that affects their final decisions. However, the use of machine learning algorithms and the large amount of structured data that is produced on a daily basis can give them a more clear and personalized view of the sentiment and trend of their portfolio assets, assisting their final decisions.

In this work, we propose ASPENDYS, an interactive web-platform that offers supportive information to investors. The primary aim of the ASPENDYS platform is to assist investors in the decision making related with stock investments, taking into consideration multiple sources and a large amount of data by using state of the art prediction algorithms. Initially, in Section 2, we review the related literature and defined the research gap and question that our work needs to cover and answer respectively. Then, in Section 3, we describe the methodology, architecture, data, components and functionalities developed in the ASPENDYS project. More specifically, the research methodology is described in Section 3.1 and the architecture of the platform is analyzed in Section 3.2. In Section 3.3, we describe all the data used by the algorithms of the components. Section 3.4 describes the algorithms that were used for the extraction of both articles sentiments and the asset sentiments. Additionally, Section 3.5 analyzes the methodology that was used in the production of the data reliability metric. Moreover Section 3.6 describes the methodologies that are supported for the optimization of the user portfolio. Finally, Section 3.7 presents the different investment signals generators that are developed and used in the ASPENDYS platform. In Section 4, we describe the resulting user interface of the ASPENDYS platform, as well as two application use cases. Additionally, in Section 5, we present the summary of this work and in Section 6 we analyze how we answered the research questions as well as the limitations of our study.

2. Related Work

In this section, a detailed literature review is presented regarding stock prediction and sentiment analysis methods. The section consists of three subsections. Initially, the related works are divided into two eras, the "before AI era" and the "AI era". Moreover, works that justify the correlation of the stock predictions and the general sentiment extracted from social media are presented. Following this, the academic gaps, the research questions that arise and the contribution of the proposed platform are defined.

2.1. Before AI Stock Predictions Solutions

The first research attempts in the field of stock prediction values are based either on econometric models or on technical indexes. The most popular statistical models, which are widely utilized in econometrics for future prices prediction, are ARIMA and ARMA [2]. Two popular indexes are utilized for stock assets price prediction: Moving Average Convergence Divergence (MACD) and Relative Strength Index (RSI). The Volatility of Average Convergence Divergence [8] is found to be effective for the prediction of a

specific asset [8]. Moreover, both indexes are proved to be more efficient in stock market trends prediction than SELL and BUY strategies [9]. Nowadays, state-of-the-art methods are based mainly on machine learning and deep learning algorithms that have been found that they achieve better results in stock market prediction.

Moreover, the stock prediction can be enhanced by, or even solely based on, market information that can be derived from the news spread and circulated in the societal sphere. Numerous studies have been conducted with the aim to capitalize on this available information, particularly since it can be accessed conveniently via digital media. Thus, the field of sentiment analysis, and the closely related opinion-mining field, emerges. The subject of the sentiment analysis is the automatic extraction of sentiments, evaluations, attitudes, emotions and opinion. The emergence and growth of these field takes place in tandem with the rise of the social media on the web, e.g., reviews, forum discussions, blogs, micro-blogs, Twitter and social networks, and is clearly related to the fact that a huge volume of data containing opinions is recorded and available. Sentiment analysis has evolved to be one of the most active research areas in natural language processing. Data mining, web mining and text mining also employ sentiment analysis. Apart from being a computer science field, it has expanded to management sciences and social sciences due to its importance to business and society and sentiment analysis systems have found their applications in almost every business and social domain [10,11]. The stock market domain of course could not constitute an exception.

Two main approaches to the problem of extracting sentiment automatically, prior to the emergence of AI, can be identified: the lexicon-based approach for automatic sentiment analysis and the statistical classifier approach. The lexicon-based approach involves calculating a sentiment for a document from the semantic orientation of words or phrases in the document. The lexicons contain words or multiword terms tagged as positive, negative or neutral (sometimes with a value reflecting the sentiment strength or intensity). Examples of such lexicons include the Hu & Liu Opinion Lexicon, the SentiWordNet Lexicon, the Multi-perspective Question Answering (MPQA) Subjectivity Lexicon, the General Inquirer, the National Research Council Canada (NRC) Word-Sentiment Association Lexicon and the Semantic Orientation Calculator (SO-CAL). The statistical text classification approach involves building classifiers from labeled instances of texts or sentences, essentially a supervised classification task [11–13]. Most notably, two lexicon-based mood tracking tools, OpinionFinder that measures positive versus negative mood and Google-Profile of Mood States that measures mood in terms of six dimensions, were used in a context related to the scope of the present work to investigate the hypothesis that public mood states are predictive of changes in DJIA closing values [14].

Issues regarding lexicon-based methods include the fact that the dictionaries are deemed to be unreliable, as they are either built automatically or hand-ranked by humans [11], as well as the need for the lexicons to be domain specific. Moreover, given a sufficiently large training corpus, a machine learning model is expected to outperform a lexicon-based model [13].

2.2. AI Era Stock Predictions Solutions

Future price prediction is a complicated task in machine learning field as it is not clear if asset prices embody investors' behavior or stock prices time series are random walks. In the literature, various AI-based and state-of-the-art studies attempted to predict future prices [3–5]. A variant of Neuro-fuzzy system which employs both recurrent network and momentum is utilized for the prediction of four assets of Dhaka stock exchange [3]. Various machine and deep learning methods are used, such as K-Nearest Neighbor regression, Generalized Regression Neural Networks (GRNN), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Gaussian Processes and Long-short term memory (LSTM) to predict future values of 111 stock market assets [4]. The employment of Wavenet model to predict the S&P 500 future prices achieves lower Mean Absolute Scaled Error (MASE) compared with the aforementioned methods [5]. The

effectiveness of Wavenet model is the main motivation to be utilized in the prediction of assets' future values in platform.

With the emergence of AI-based techniques, deep learning neural networks became the dominant approach to tackle natural language and text processing. These deep learning approaches include many networks types such as Fuzzy Neural Networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and more recently transformer-based neural networks [15–17]. Numerous studies have been conducted aiming to leverage the great amounts of available data to extract more accurate and exploitable stock market information. Sentiment analysis and machine learning principles were applied to find the correlation between "public sentiment" and "market sentiment"; firstly, Twitter data were used to predict public mood and subsequently stock market movement predictions based on the predicted mood and the previous days' DJIA values were performed [15]. In a later work, sentiments were derived from Yahoo! Finance Message Board texts related to a specific topic of a company (product, service, dividend, etc), employing thus the "topic sentiment" and not the "public sentiment". The sentiments were incorporated into the stock prediction model along with historical prices [18]. Sentiment analysis and supervised machine learning principles were also applied to the tweets about a company and the correlation between stock market movements of the company and the sentiments in tweets was analyzed, concluding that a strong correlation exists between the rise and fall in stock prices with the sentiments expressed in the tweets [19]. In [6], the proposed prediction model was based on sentiment analysis of financial news and historical stock market prices. In the scope of the specification of trading strategies, taking into account the decaying effect of news sentiment, thus deriving the impact of aggregated news events for a given asset, was also proposed [20]. Trading strategies that utilize textual news along with the historic prices in the form of the momentum to obtain profits were designed by Feuerriegel and Prendinger [7].

Many works have been published concerning financial forecasting and portfolio management platforms utilizing AI technologies. Researchers are addressing the challenges arising from the application of AI technologies in quantitative investment. In particular, they have designed and developed an AI-oriented Quantitative Investment Platform, called Qlib. It is based on attempts to build a research workflow for quantitative researchers, utilizing the potential of AI technologies. Their platform integrates a high-performance data infrastructure with machine learning tools dedicated for quantitative investment scenarios. It provides flexible research workflow modules with pre-defined implementation choices regarding investment, as well as a database dedicated to scientific processing of financial data, typical for tasks in quantitative investment research. Finally, by including basic machine learning models with some hyper-parameter optimization tools, the authors reported that Qlib outperforms many of the existing solutions [21].

There are recommendation systems supporting investment decisions, which are based on extracting buy/sell signals retrieved from technical analyses and predictions of machine learning algorithms. Through the platform's interface, users are able to derive their own conclusions and manage their investment decisions by evaluating the analyzed information by the system. The authors' development was based on retrieving historical financial data regarding various financial products: stocks, funds, government bonds, certificates, etc. The predictions on the future behavior of a product were derived through machine learning regression algorithms, such as Random Forest, Gradient Boosting, MLP and KNNeighbors. The technical analysis factors were calculated by utilizing indicators such as Relative Strength Index (RSI), Stochastic Oscillator (STOCH), Simple Moving Average (SMA), Exponential Moving Average (EMA) and others. Finally, their platform provides representations of prices' time-series, tables and charts including financial products' open-high-low-close values that depend on the aforementioned analyses [22].

Some works integrate sentiment analysis with a SVM-based machine learning method for forecasting stock market trends [23]. Their contribution consists of considering the day-of-week effect to derive more reliable sentiment indexes. The financial anomaly day-

of-week effect [24], which indicates that the average return on Mondays is much lower than that on the other weekdays, influences the accuracy of the investor sentiment indexes. Researchers are using this effect by introducing an exponential function on past sentiment changes on weekends and then generalize to holidays, to adjust the sentiment indexes. To forecast stock market movement direction, they extract features from unstructured financial review textual documents from which the sentiment indexes are constructed. Lastly, a SVM model is employed to produce the stock price predictions by implementing five-fold cross validation and a rolling window approach.

Some works employed sentiment analysis on Twitter posts in an attempt to produce social signals to be used along with historical market data, in order to predict cryptocurrency prices. They tackled the sentiment analysis task using VADER, and, along with features produced from market data, they trained MLP, SVM and RF algorithms [25].

There are works that used stock price historical data in combination with financial news articles' information to predict the prices of stocks. Testing the performance of ARIMA, Facebook Prophet and RNN models without textual information, as well as the performance of RNNs with additional textual information, they concluded that the latter performed better. More precisely, their results show that an RNN using price information combined with textual polarity calculated by the nltk Python library, in most cases performed better than the aforementioned models as well as than that of an RNN operating with price data along with the text as inputs [26].

2.3. Defining the Academic Gaps and Research Questions

Based on the aforementioned analysis, it is concluded that there are various methods in the literature that deal with the stock prediction. These methods utilize either statistical methods or AI methods including DNNs. Most of the presented methods use historical values of the stock prices in order to predict the forthcoming prices, while there are others that try to predict stock prices through sensing the general sentiment on social media.

Even though there are many works dealing with stock prediction, there are still some gaps that have to be considered. First, for the stock prediction, most of the existing works are based on stocks' historical values. This approach lacks of considering other information that may be significant, such as social or political events that are depicted in news articles and social media posts. Such information may lead to the early detection of critical changes in the trends of stock prices. Feuerriegel and Prendinger [7] proved that such information can be considered by sensing public sentiment on social media. Thus, the monitoring and analysis of social medias' sentiment can be of added value for the stock prediction. In this aspect, there is a lack of platforms that handle and combine both historical stock prices and the general sentiment for the investment recommendation. Second, although there are many platforms or systems that provide advice for investments, also called robo-advisors, they usually act automatically without providing to the investors the information that led to their recommendations. Thus, the interpretability of their recommendations is not well established. Finally, there is the need for an interactive platform that will collect and analyze data in a daily basis and provide investment recommendations not only for specific assets but for an investor's portfolio. The suggestions should be personalized to the profile of the users so that different advice can be provided to risky and conservative investors.

Following the above research gaps, we propose ASPENDYS, an interactive web-platform that offers supportive information to investors, thus facilitating the decision making related with stock investments. The platform consists of several tools and methods that utilize key technologies of the computer science. These include Natural Language Processing (NLP), a sub-field of computer science, and artificial intelligence that can accurately extract information and insights contained in the textual data as well as categorize and organize these data. Additionally, in the ASPENDYS platform, deep learning (DL) technologies were used, which is a sub-field of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

The research questions (RQ) that have arisen are as follows:

RQ1: Is the ASPENDYS platform able to provide investment recommendations by combining both historical stock prices and the extracted sentiment, for an asset, through the sensing of social media and news websites?

RQ2: Can ASPENDYS platform be used as a supportive tool for investors that will assist their investment decision making and monitoring?

3. Materials and Methods

3.1. Methodology

Initially, the work started with the collection of the requirements from real users. For this purpose, a questionnaire was created that was shared to 30 individuals of the investment private sector in Greece. The end users became aware of the scope of the ASPENDYS project and the goals of the systems that is going to be developed before filling in the questionnaire. The results of this research led to the definition of the specification of the ASPENDYS platform as well as its architecture. More specifically, we concluded that there was a necessity for implementing components that will collect data from different sources, check the reliability of these data, generate investment signals and optimize the user portfolio. All of these components should be integrated into a web application with a friendly user interface. Following that, we defined the system architecture and declared the connections among the platform components as it is described in the following sections. The next step in our research was the literature review for the specific research field in order to identify the state of the art in the stock market prediction, as described in detail in the previous section. Based of this literature review, we started implementing each of the architecture components. Moreover, we defined the sources of both textual and historical stock prices. Regarding the textual data, we defined that we should use news websites in order to collect news articles that regard certain assets. In addition, for the collection of information from social media, Twitter and Stocktwits were selected. Twitter is one of the most popular social media sites where users express their opinion on various topics. In addition, they utilize the functionality of hashtags, which allows the filtering and retrieval of information in an efficient way. Moreover, Stocktwits was selected, as it is very similar to Twitter and its users are mainly investors. Regarding the retrieval of the stock prices, Yahoo! Finance was selected, since it provides both historical and up-to-date stock prices that are updated daily. The final step of our work was the integration of the modules to a unique platform that exports each component service through RESTful APIs to the ASPENDYS web application as well as to run several use cases in order to validate our system, which is described in Section 4.

3.2. System Architecture

In Figure 1, the architecture of the ASPENDYS platform is illustrated. The platform is separated into seven modules: Data Collection, Database Management System, Portfolio Optimization, Sentiment Analysis, User Portfolio Management and User Interface.

The Data Collection module is responsible for the extraction of data from various data sources. It consists of two sub-modules: one retrieves data from the social media using the Twitter and Stocktwits API, and the other extracts articles and financial data using the News API and the Yahoo! Finance API. The Database Management System is responsible for storing and retrieving the data from news articles and financial indicators that are collected from the Data Collection module. Additionally, this module is used for user management, for both registration and user data retrieval. The DBMS that is used is a MySQL server, and the application logic is implemented in Python. The Portfolio Optimization module is triggered by the User Interface, and its results are shown on it. This module proposes a better synthesis of the portfolio based on different methods, which are analyzed in the following subsection.

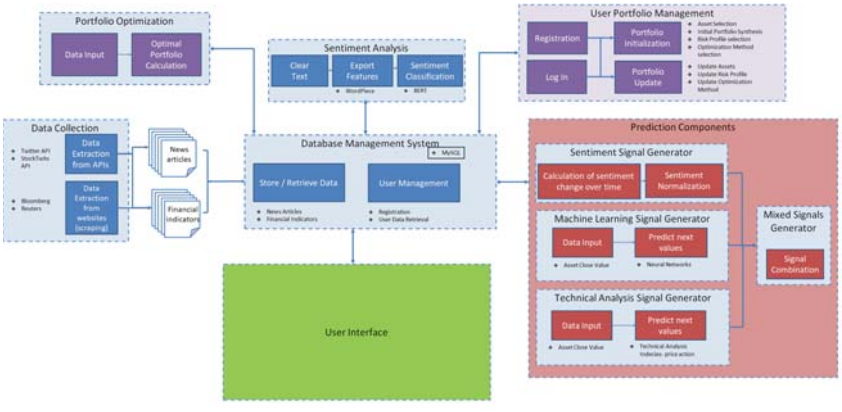


Figure 1. ASPENDYS platform architecture.

The Sentiment Analysis module uses input from the Data Collection module through the database in order to produce the sentiment of the articles and the respective assets. Using the User Portfolio Management module, the user of the platform is able to create a new portfolio as well as update the current one. The interaction with this module can be done through the User Interface and the results are stored in the Database Management System. The Signal Generators module consists of four sub-modules that generate investment signals using different input data. These sub-modules are sentiment, machine learning, technical analysis and mixed signal generators; more details are provided in the following sections. The User Interface is the front-end of the ASPENDYS platform where the user is able to visualize the data in graphs and tables, as well as manage their portfolio. It is an Angular application [27] that communicates with the data management system as well as the portfolio optimization and the user portfolio management module.

3.3. Data Collection

In this section, the data collection processing of both texts and prices is described. Texts are sorted in short-text data derived from Twitter and Stocktwits and long-text data derived from financial news article platforms, whereas stock prices come from the Yahoo! Finance site. All data, both textual and numerical, are stored and integrated in SQL type database. In next subsections, data collection processing and data features are presented in detail.

3.3.1. Twitter

To collect tweets from the social media platform Twitter, we used tweepy [28], an open-source Python library, which handles the connection to Twitter's API [29] given the required account information and streams tweets based on certain word filters and other search parameters. On a daily basis, the platform collects tweets for each asset, based on the condition of including the asset's name or using a set of relevant hashtags. The information stored for each tweet is tweet's id, text, number of re-tweets, creation date, user's screen name, whether the user is verified, user creation date, number of followers, status count, number of friends and, finally, number of mentions, URLs and hashtags used. A tweet is stored in the database, after checking its credibility and if its id (incremental integer value) is larger than the most recently stored tweet in the database, ensuring that no duplicates are stored. The credibility of the tweets is checked by classifying them into two classes (credible or non-credible) by a Random Forest classifier trained on the CREDBANK dataset [30] and based on the rationale described in [31]. The features of tweets used as input for the classifier are: the user creation date, whether the user is verified, the status count and the numbers of followers, friends, mentions and URLs and hashtags used.

3.3.2. Stocktwits

Regarding the collection of data from Stocktwits social media platform, we used Stocktwits' API [32]. Stocktwits for each asset are retrieved daily using the asset's ticker. The information stored for each one is the id, text, number of mentioned users, creation date, sentiment of the stocktwit as generated by the platform, username, user creation date (join date), whether the account is official, number of followers and number of accounts that the user is following. Consequently, the credibility of the stocktwits is checked, and they are stored in the database if their id is larger, meaning they are more recent than the last Stocktwits post in the database. The credibility of the stocktwits is checked using the same Random Forest classifier used in the case of tweets retrieved from Twitter, the only difference being the features used as input for the classifier. Since less information is available for these stocktwits, only the creation data, whether the account is official and the numbers of followers and accounts that the user is following are used as input for the classifier, with all other count features regarded as equal to zero.

3.3.3. News Articles

To perform sentiment analysis on news data, concerning the involved assets, we collect news articles on a daily basis from six liable sources: BBC News, Reuters, Coin-telegraph, Yahoo! Finance, The Wall Street Journal and Bloomberg. The search terms applied to retrieve relevant articles is the asset's name, as well as the asset's ticker. For some assets, only the asset's name was used as a search term, as their ticker is a too vague term (e.g., C for Citigroup, V for Visa or F for Ford). For the articles' collection, a set of web scrapers were developed and used in combination with the Python library NewsAPI [33].

The following information is collected for every gathered news article: the article's title, the full text, posting date, writer and source's website. Subsequently, the collected articles are checked to not have a very similar body with each other, as well as with every article in the database. To achieve this, the text similarity between each pair of articles is calculated, utilizing the NLP Python library Spacy [34]. The motivation behind this is to avoid storing and analyzing the same article multiple times, as particularly similar articles could indicate that the same article has been retrieved multiple times from different sources or that the same article with minor changes has been re-posted. In both cases, very similar articles need to be filtered out.

3.3.4. Assets Values

For the collection of the assets values, the Python module "pandas datareader" [35] was used, in order to extract data from the Yahoo! Finance API. More specifically, it collects open, close, high, low, volume and adjusted close values for each asset on a daily basis. The open value is the initial price of each asset when a trading session starts and the close value is the last exchange price at the expiration of a trading session. The adjusted close value amends an asset's close price to reflect that asset's value after accounting for any corporate actions. The high value is the highest selling price of the asset and the low value is the lowest selling price of the asset during trading hours. The volume value is the total number of shares exchanged between stakeholders during trading hours multiplied by the current selling price of the asset.

3.4. Sentiment Analysis

In this paper, we apply text sentiment analysis to extract the sentiment of stock-related texts, namely articles, tweets and stocktwits. We aim to capture the sentiment regarding a specific asset as expressed in a number of texts whose reliability we have previously confirmed.

3.4.1. Articles and Tweets Sentiment Method

We perform sentiment classification per article, tweet or stocktwit on 5-point scale ranging from -2 to 2. The scores $[-2, -1, 0, 1, 2]$ correspond to the five different classes

[very negative, negative, neutral, positive, very positive]. (Through the UI, the user has access to the sentiment value attributed to each article, tweet or stocktwit, as well as the text of the article itself).

For the classification task, we used the neural network transformer-based model, Bidirectional Encoder Representations from Transformers (BERT) [17]. BERT is a language model for Natural Language Processing (NLP) that has achieved state-of-the-art results in a wide variety of NLP tasks. To adapt BERT to our specific task, we fine-tuned it on the Stanford Sentiment Treebank-5 (SST-5) dataset [36] for fine-grained sentiment classification. The fine-tuned model accepts as input the text and outputs the class ([−2, −1, 0, 1, 2]) attributed to text providing thus an estimation of the sentiment best describing the text.

3.4.2. Asset Sentiment Method

We approach the task of estimating the total sentiment corresponding to a specific asset in the present moment by equally weighting the asset sentiment derived from the most recent, newly sourced, texts and the averaged sum of the previous decayed asset sentiment values and computing their average. For the computation of the asset sentiment from the recent texts, we first aggregate all the recently acquired and not processed yet texts (articles, tweets and stocktwits) related to the specific asset, and then, using the model described above, we attribute sentiment values to each of the texts and after that we average the values.

3.5. Author and Source Reliability

Author and source reliability is the process where an indication of the level of trustworthiness of an article author or of a news feed site is extracted. This module is not related with social media platform post reliability but it does correspond with news media articles reliability. Unless news media articles are considered legit and reliable, ASPENDYS platform receives numerous articles from various sources so the extraction of their reliability is useful information. This task is complicated because there are no objective criteria for the reliability of an author or a media source according to recent studies. More specifically, the annotation of news articles is required to be performed by humans to classify them as either reliable or unreliable [37,38]. This process has been found to be biased since humans proved to be affected by the popularity of the author or news site [38,39]. Additionally, in some cases, it is taken for granted that textual data are labeled as written by a reliable author or not [40]. It is remarkable that state-of-the-art methods focus either on fake news detection or on author reliability rather than on the employment of a global method resolving both of these issues. In this study, based on the intuition that reliable level is not absolute but relative, both author and source reliability are estimated based on similarity with the rest of all of the items. More specifically, a profile for each author and source is generated and for each these items the average similarity with the others is computed; the item with the highest similarity is assigned a predefined high score; and, lastly, for the rest of items, values proportional to the highest value are assigned. Generally, in the case of author and source reliability, the more similar a profile is to the others, the higher is the reliability score assigned to it. To estimate the similarity between items, we utilize the function words for each sentence [41]. Function words were found to be content-independent and efficient in the extraction of writing profile of authors. More specifically, they belong to a specific set of parts of speech such as prepositions, conjunctions and pronouns. Each function word is relative to either itself or another function word if they are inside the same sentence provided they are contained in the processing window. The assigned score for each word is based on both adjacency network and Markov chain model, as its relation with another word decreases as their distance increases. For the comparison of the writing profiles of authors and sources, Kullback–Leibler divergence is utilized.

3.6. Portfolio Optimization

Portfolio Optimization is the process that determines the distribution of stock market products that make up a portfolio based on certain sizes. These figures are usually the expected profit and risk, always based on the assumption that the risk is intertwined and proportional to the profit. The real revolution in portfolio management came in 1952 from Harry Markowitz (Nobel Prize in Economics in 1990), who introduced and built what became known as Modern Portfolio Theory (MTP). Modern Portfolio Theory is aimed at investors. It applies the way in which we can optimize/maximize the expected return of a portfolio given the level of risk we want to take. Similarly, given the desired performance of a portfolio, we can build a portfolio with the least possible risk [42,43]. Regarding the current project and for the Portfolio Optimization of the users, we applied the follow methodologies:

- Portfolio Optimization based on Modern Portfolio Theory [44] is based on fundamental concepts of Statistics such as Variance and Correlation. It proves that it is more important to measure the performance of a portfolio by the total portfolio of products that make it up. The set of Optimized Portfolios constitutes a curve called the Efficient Frontier. Thus, with each desired expected return (Y-Return axis), we are able to know our Optimized Portfolio that results from the corresponding point in the curve and of course corresponds to the minimum risk. Similarly, for each level of risk we are willing to take (X axis—Standard Deviation (Risk)), we look for the corresponding point in the Effective Front curve, which returns us the Optimized Portfolio with the maximum possible return.
- Portfolio Optimization based on Modern Portfolio Theory is the selection of the minimum variance-risk (Minimum Variance Portfolio), which is a variation of the Modern Portfolio Theory. From the Efficient Frontier curve, we choose the point that constitutes a diversified Optimized Portfolio consisting of stock products that provide the least possible risk.
- Portfolio Optimization Based on the Black–Litterman Model is a mathematical Portfolio Optimization model developed in 1990 at Goldman Sachs by Fischer Black and Robert Litterman and published in a more enriched version in 1992 in the Financial Analysts Journal. This model solves some problems faced by institutional investors but also ordinary traders when they apply the Modern Portfolio Theory in practice. The innovation of Black–Litterman model is that the investors are able to define certain views for each one of the assets of the portfolio [45–47].
- Portfolio Optimization based on the Risk Parity model is a Portfolio Optimization strategy used extensively by many investment schemes (mainly hedge funds) for maximum return, given a level of risk through the portfolio risk method (Equally-weighted risk contributions portfolio). Each product participates in the Portfolio in the same way as in the total variation of the Portfolio [48,49].

3.7. Signals Generators

In the proposed platform, we developed four different investment signals generators. More specifically, the generator based on the asset sentiment, the machine learning generator, the technical analysis generator and the mixed generator. These generators produce two types of signals: the SELL signal that proposes to the user to sell the asset and the BUY signal that proposes to user either to increase the percentage of investment in a specific asset or to buy this asset if it is not included in their portfolio. In the following subsections, we describe in details these components.

3.7.1. Sentiment Signal Generator

The method (strategy) we implemented for the sentiment signal generation utilizes the (total) asset sentiment along with the historic prices in the form of the momentum. The rationale behind this choice is that we want to invest in assets with both a high polarity sentiment and previous momentum in the same direction. Thus, trading signals based on sentiment are generated only if both the asset sentiment and the asset's historic prices

provide an evidence of continuation in the same direction. Regarding the sentiment analysis expressed in the articles and the social media posts, the fact that there was no available dataset containing financial texts annotated with fine-grained sentiment labels constituted a challenge. For this reason, the Stanford Sentiment Treebank-5 (SST-5) that contains reviews from the Internet Movie Database (IMdb) was used to fine-tune our model, based on the fact that is constitutes a reliable dataset with fine-grained (5 classes) sentiment annotations.

3.7.2. Machine Learning Signals

Financial Signals are based on predicted close values, which are formed to belong to one of the following categories: BUY and SELL. For the extraction of the predicted values, a deep learning method that is based on a modified version of WaveNet model [50] was found to be effective in the multivariate forecasting of financial time series [5].

The prediction of time series is implemented under the condition of other highly correlated time series based on Wavenet model standards. The WaveNet model [50] is implemented with the application of a dilated convolution neural network to distinguish the predicted prices from the actual prices, instead of utilizing the recurrent connections in which the preceding values are taken into account. In the dilated convolution process, a filter is applied in a more extended area than its length and the input values are skipped by a stable step.

The architecture of the method combines one convolutional layer and two dense layers with Relu activation function. The hyper-parameters of the model were set as follows: the number of kernels equal to 64, the kernel size equal to 2, the time-step equal to 1 and the processing window equal to 64. The input vector is based on 60 financial indices found to be effective in stock price forecasting [51] and is estimated by the combination of open, close, low, high, volume and adjusted close values of market stocks. Input data are initially preprocessed by applying the min-max normalization method on them.

Feature selection is based on method developed by Yuan et al. [52]. More specifically, the Random Forest feature selection algorithm is applied on the input data. A financial index considered more important when it increases the difference between the out of bag score in a feature vector with added random noise and the out of bag score in the initial feature vector. Finally, 44 financial indices out of 60 are selected. The post-processing of predicted values is defined as their transformation to financial signals, so each prediction should be mapped to one of the two aforementioned categories. For this purpose, each prediction is transformed to its percentage variation in relation with the preceding prediction value to be estimated as the return values of each asset. In the case of the return value is higher than double the summation of average and standard deviation of the return distribution, the extracted signal is equal to BUY. When the return value is lower than double the standard deviation minus the average of return distribution, the signal is assigned to be equal to SELL.

3.7.3. Technical Analysis

Technical Analysis is an investment and trading tool that spots investing opportunities and produces trading signals by analyzing statistics that come from the traders and markets' activity such as the price action and the trading volume. Technical Analysis focuses mainly on the study of the price and the volume of a trading asset.Additionally, the Technical Analysis is based on the Efficient Market Hypothesis, a hypothesis that states that the prices of stock market products reflect all the information that concerns them. Technical analysis also assumes that values move in iterative patterns and that these patterns are often repeated. The main problem that Technical Analysis faces is that it often contains subjectivity. Indicators, oscillators and other objective tools should be used; all technical analysts should see the same indicator and the same values, without being dependent on them. The portfolio optimization presents various challenges such as the fact that too many transactions are proposed. This is addressed by properly configuring the system and selecting low-cost changes. Minimal variation portfolios without the expected returns can

also be proposed. This can be resolved by minimizing the tracking error in the portfolio. At the same time, the mean variance optimization is restrictive. In this case, an optimization algorithm can be selected with a few limitations. As many times one of the products participating in the portfolio takes too much weight in the portfolio, , the selection of products should be based on specific rules to avoid this. The tools of Technical Analysis thoroughly study the ways of supply and demand of a trading asset and how these factors affect the price, trading volume and volatility. Technical Analysis is also used for the production of short-term trading signals through charting tools as well as a tool for the evaluation of certain trading assets regarding the relative markets or relative sectors [53,54]. Recently, many new developments have appeared in the Technical Analysis field including methodologies, algorithms, indicators, oscillators, etc. [55], but they are all based on three very simple assumptions [56]:

1. Prices reflect anything: Most technical analysts believe that any factor that is correlated to a trading asset, which is tradable in the global markets, from the fundamentals up to the market psychology, is already reflected in its price. This assumption is aligned with the famous Efficient Market Hypothesis (EMH), which comes to the same conclusions regarding the prices. Therefore, what technical analysts suggest is price analysis, which, according to the aforementioned assumption, contains all of the information of the asset such as the supply and demand of a stock or a commodity.
2. Price movements are trending: Technical analysts expect that the asset prices, even in random market movements, come out in trends, no matter what the observation timeframe is. The price of a stock, most likely, will continue following the trend, instead of moving randomly or erratically. Most Technical Analysis strategies contain trend as a basic feature.
3. History tends to repeat itself: Technical analysts believe that history repeats itself. The repetition of price movements is often attributed to market psychology that tends to be quite predictable as it is based on emotions such as fear, over-optimism, panic, etc. Technical analysts use chart patterns to analyze emotions and possible subsequent movements in the markets to understand trends. In recent decades, countless trend-setting techniques have been developed, with different approaches and tools.

Regarding the current project, to extract trading signals, we used the following Technical Analysis tools and indicators:

- MACD Indicator (Moving Average Convergence-Divergence) to define the market trend.
- ADX Indicator (Average Directional Index) to define the market trend strength.
- High and low prices in specific timeframes to find supports and resistances (price levels that are difficult to breakout either in a bullish trend (resistance) or in a bearish trend (support).
- RSI Indicator (Relative Strength Index) to define any possible overbought/oversold price levels in order to spot any possible corrective movements or even a trend reversal.
- ATR Indicator (Average True Range) to define the market volatility in order to apply the optimum risk management as well as the stop loss and take profit levels.

Trading signals (BUY/SELL) are produced, by combining all the above indicators [53–58]:

- MACD Indicator gives the market trend. MACD > 0 means uptrend, which is a BUY signal, while MACD < 0 means downtrend, which is a SELL signal.
- ADX Indicator should be >25 in order to have enough market strength. It is a filter to both BUY/SELL signals.
- Current price should NOT be close to resistance for BUY signals or close to support for SELL signals. It is a filter for both BUY/SELL signals.
- The asset should NOT be in overbought condition for BUY signals or in oversold condition for SELL signals, as defined through RSI indicator.
- The suggested size of each signal is defined through ATR as well as the suggested take profit/stop loss levels.

3.7.4. Mixed Signal Generator

Machine learning, Sentiment Analysis and Technical Analysis modules consider each asset price as independent. In mixed signals module, the intuition that each asset behavior is affected by prices of other assets is taken into account. Signals of all modules are integrated and it is assumed that each asset close price prediction should be based on both close prices and signals of itself and others.

For this purpose, it is essential for close prices of assets to be transformed in financial signals. Based on the moving average rule described in work [59], the extracted mixed signals are compared to the values between long-run and short-run moving averages with a band value taken into account. More specifically, in the case that the long-run moving average is higher than the summation between the corresponding short-run and the band value, the BUY signal appears.

In contrast, a SELL signal is derived in the case where the long-run average is lower than the short-run counterpart minus the band value. Cases in which the difference between the long-run and short-run moving average is between 0 and the band value are not taken into consideration. The long-run moving average is defined as equal to 50, the corresponding short-run equal to 1 and the band equal to 0.01 as dictated by Gunasekarage and Power [59] as the optimal parameters combination to achieve maximum profit. To train the input data, the method described by Spyromitros-Xioufis et al. [60] is employed, which is an advanced version of logistic regression algorithm with the extension of a Lasso regularization term.

4. Results

The results of this research concern the presentation of the user interface functionality as well as some use cases of the presented platform.

4.1. Aspendys User Interface

The ASPENDYS User Interface is separated into two main views that contain widgets: the Portfolio Management view and the Investment Sentiment Analysis view. Both views share a common component which is the Investment Signals. In the following subsections, these resulting views are analyzed in detail.

4.1.1. Portfolio Management

The Portfolio Management view is one of the two main views of the ASPENDYS platform. It contains four unique widgets that assist the user to be informed about the process of their portfolio as well as to manage it: (1) the Portfolio Overview; (2) the Profit/Loss History; (3) the Signal Acceptance History; and (4) the Current Portfolio Synthesis. These four widgets are included in the specific view, as shown in Figure 2 and analyzed in this section. The Portfolio Management view also contains the widget Investment Signals that is common in both views of the platform.

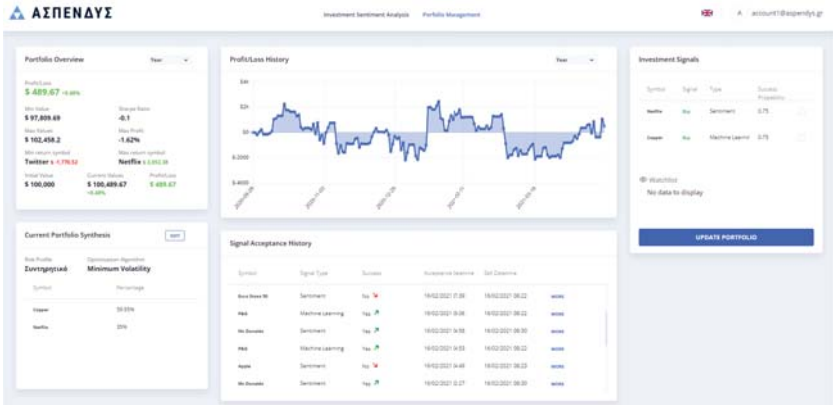


Figure 2. ASPENDYS Portfolio Management.

- The Portfolio Overview widget contains information regarding the user portfolio and their values are updated each time the platform receives news asset values. Specifically, the information illustrated contains:
  - Profit/Loss: The profit or the loss of the portfolio in both absolute value (dollars) and percentage.
  - Min Value: The minimum value in dollars that the portfolio has reached in the selected date period.
  - Max Value: The maximum value in dollars that the portfolio has reached in the selected date period.
  - Sharpe Ratio: The Sharpe Ratio value of the portfolio.
  - Min Return Symbol: The symbol with the minimum return as well as the absolute value in dollars of the return.
  - Max Return Symbol: The symbol with the maximum return as well as the absolute value in dollars of the return.
  - Initial Value: The initial investment value; by default, we have set this value to $100,000 for each user.
  - Current Value: The current value of the portfolio, which is the addition of both cash and assets value.
- The Profit/Loss History widget contains a line chart with the time series of profit/loss in the selected time period. The available periods are the current week, month and year.
- The Signal Acceptance History widget contains a table with all the BUY signals that have been accepted by the user. This table consists of:
  - Symbol: The name of the symbol to which the signal refers.
  - Type: The type of the signal depending on the generator that produced it (sentiment, technical, machine learning or mixed)
  - Success: A Boolean value that declares if the portfolio gain value is following this signal or not.
  - Acceptance Date-Time: The date and time that the user accepted this BUY signal for the specific asset.
  - Sell Date-Time: The date and time that the user accepted a SELL signal for the specific asset.

- The Current Portfolio Synthesis widget contains the percentages of all the assets of the portfolio as well as the selected Risk Profile and the Optimization Method. Moreover, in this widget, the user is able to edit their portfolio, as shown in Figure 3, by adding new assets, increasing their investment percentages or selling assets. Additionally, the user is able to change the Risk Profile or the Optimization Method, which are inputs to the portfolio optimization service.

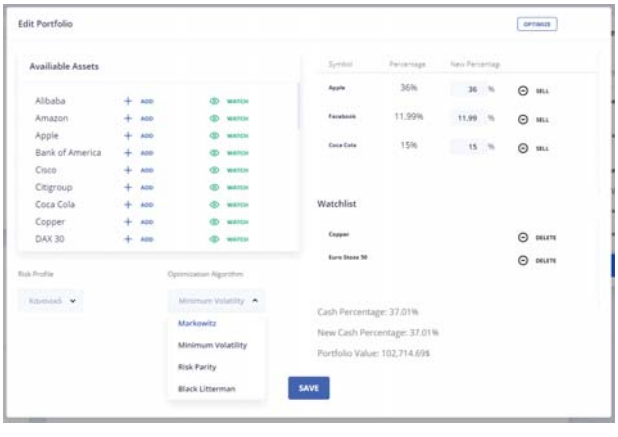


Figure 3. Edit User Portfolio.

4.1.2. Investment Sentiment Analysis

The Investment Sentiment Analysis view is one of the two main views of the ASPENDYS platform. It contains unique widgets that assist the user to be informed about the sentiment of their portfolio assets, as well as the sentiment of articles that refer to these assets. More specifically, the Portfolio Investment Sentiment, the Investment Sentiment Change Notifications and the Sentiment Analysis of News and Articles are the three widgets that are included in the specific view, as shown in Figure 4 and analyzed below. The Investment Sentiment Analysis view also contains the widget Investment Signals that is common in both views of the platform. The specific view is fed by data that are generated by the components analyzed in Section 3.4.

Figure 4. ASPENDYS Investment Sentiment Analysis.

- The Portfolio Investment Sentiment widget contains a pie chart with the percentages per sentiment for all the portfolio assets as well as the absolute value in dollars that corresponds to each sentiment. Moreover, this widget contains the current portfolio value as well as the percentage of profit or loss respectively. Additionally, in this widget, all the portfolio assets are displayed along with the state of their sentiment and the probability that this sentiment is accurate.
- The Investment Sentiment Change Notifications widget contains a table with all sentiment changes of the portfolio assets. The columns of this table are:
  - Symbol: The name of the symbol to which this entry refers.
  - Sentiment: The sentiment status of the specific symbol. The values it can be assigned are Very Negative, Negative, Neutral, Positive and Very Positive.
  - Change: The combination of the previous with the current state of sentiment.
  - Index: The index shows how much a symbol's sentiment state has improved or worsened.
  - Date/Time: The date and time describe when this change was detected.

  By selecting one of these changes, the user is able to see the timeline of the asset values together with the produced signals from the components in Section 3.7, as shown in Figure 5.

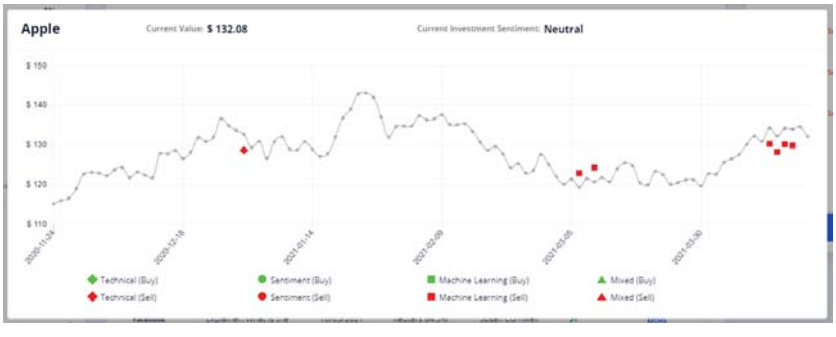


Figure 5. Asset values and signals timeline.

- The Sentiment Analysis of News and Articles widget contains the most recent articles and news that correspond to the assets of the users portfolio. The sentiment of each article is generated by the components analyzed in Section 3.4.1.

4.1.3. Investment Signals

The investment signals widget is a common widget in both Portfolio Management View and Investment Sentiment Analysis View. The data used in this widget are generated by the components in Section 3.7. As shown in Figure 6, it contains a table with the most recently generated signals for the assets that belong to the user portfolio.

The columns of the investment signals table include the symbol, signal, type and success probability, analytically:

- Symbol: The name of the asset that the signal refers to.
- Signal: The signal value can be either SELL or BUY.
- Type: The type of signal that can be Technical, Sentiment, Machine Learning or Mixed and indicates the generator that produced this signals Technical Analysis Generator, Sentiment Analysis Generator, Machine Learning Generator and Mixed Signals Generator, respectively.
- Success Probability: This probability is a metric that generated by each component and indicates the probability of this signal being successful.

The user is able to select one or more signals that they want to accept, and, by pressing the update portfolio button shown in Figure 6, the view in Figure 7 appears, where the

user is able to increase the percentage of the assets that have BUY signals or sell the assets with SELL signals.

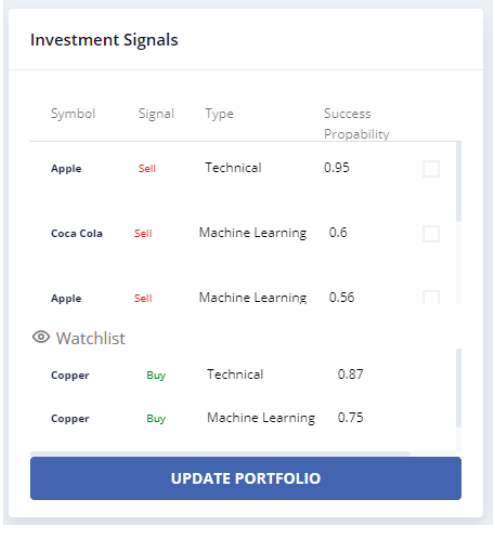


Figure 6. ASPENDYS Investment Signals.

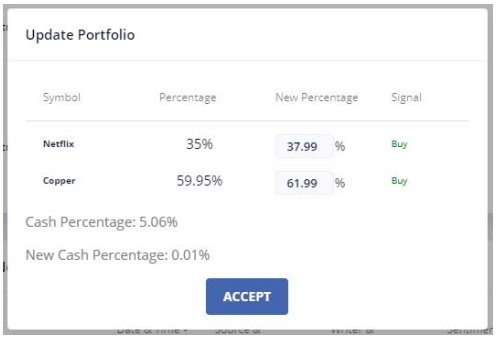


Figure 7. Update Portfolio using signals.

4.2. Application Use Cases

In this section, we analyze two different use cases of the ASPENDYS platform. The first use case is about the portfolio of a user that has invested a large amount of money into specific assets that have large value per piece using an aggressive risk profile, while the second use case is about a portfolio that has invested a smaller amount of money to assets with smaller value using a conservative risk profile.

More specifically, in Table 1, the assets of the aggressive portfolio are depicted as well as the amount that has been invested in each of them when they were initialized on 01/02/2021. The total amount of the investment is $79,833.25 and the remaining cash is $20,166.75.

Regarding the aggressive portfolio, on 05/02/2021, the sentiment analysis generator produced a BUY signal for the asset Twitter and the investor of this portfolio accepted this signal increasing the percentage of this asset to 35.22%. On 16/02/2021, the Technical Analysis Generator produced a SELL signal for the assets Facebook and Alibaba and the investor of this portfolio accepted this signal. Additionally, on 01/03/2021, the Mixed Generator produced a SELL signal for the asset Netflix and the investor of this portfolio

accepted this signal. The result of this use case was to increase the value of the portfolio by 2.95%, while, if the investor did not accept any of the signals, the value of the portfolio would increase by 0.69%. Figure 8 shows the progression of the portfolio value. The green line shows the actual value of the portfolio with the signals that have been accepted, while the blue line shows the portfolio value if the signals had been discarded.

Table 1. The assets of the aggressive portfolio, together with the invested amount in dollars as well as the percentage that each asset takes up in the portfolio.

Asset	Percentage	Amount
Facebook	22.97%	$22,974.64
Alibaba	18.99%	$18,994
Twitter	16.89%	$16,891.3
Netflix	20.97%	$20,973.24

Figure 8. Aggressive portfolio value timeline.

In Table 2, the assets of the conservative portfolio are illustrated as well as the amount that has been invested in each of them when they were initialized on 01/02/2021. The total amount of the investment is $26,534.11 and the remained cash $73,465.89.

Table 2. The assets of the conservative portfolio, together with the invested amount in dollars as well as the percentage that each asset takes up in portfolio.

Asset	Percentage	Amount
Google	5.86%	$5863.11
Apple	6.11%	$6119.84
Oracle	4.25%	$4259.61
Tesla	8.09%	$8091.89
Cisco	2.19%	$2199.66

Regarding the conservative portfolio, on 06/02/2021, the Machine Learning Generator produced a SELL signal for the asset Apple and the sentiment analysis generator produced a SELL signal for the asset Tesla, the investor of this portfolio accepted these signals. On 22/02/2021, the Technical Analysis Generator produced a BUY signal for the asset Oracle and the investor of this portfolio accepted this signal increasing the percentage of this asset to 18.32%. The result of this use case was to increase the value of the portfolio by 3.28%, while, if the investor did not accept any of the signals, the value of the portfolio would have decreased by 2.14%. Figure 9 shows the progression of the portfolio value. The green line shows the actual value of the portfolio with the signals that have been accepted, while the red line shows the portfolio value if the signals had been discarded.

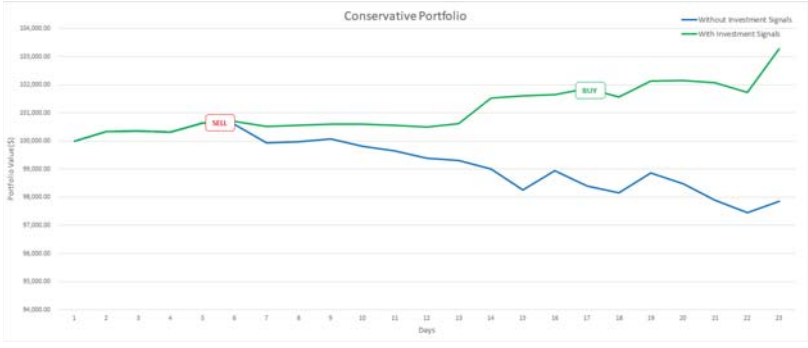


Figure 9. Conservative portfolio value timeline.

5. Discussion

In recent years, the revolution in data analysis offered by the big data approach enables the generation of new AI tools in several fields. These advances have created the need for new decision-making systems for the generation of modern investment models. The current trends in the financial sector create the demand for automated portfolio management services. This type of portfolio management can combine several approaches such as machine learning using data analysis to select underlying investments and technical analysis systems assessing the likely magnitude of potential price moves based on historical correlations. Moreover, nowadays, opinions are no longer something that needs to be sampled via focus groups or surveys. Instead, content derived from social, event and news media sites and vendors who interpret and spread those data are providing data streams that offer insight into the thinking of their myriads users. The sentiment analysis of news stories and social posts using machine learning is a trending topic. There are trained models that automatically identify if social media posts and news stories about an asset are positive or negative, assisting investors to stay ahead of future price movements.

In this paper, we describe in detail the ASPENDYS platform, an AI-enabled platform for financial management. More specifically, the ASPENDYS platform is a high-level interactive interface offering end users the functionalities of monitoring, modifying and expanding their portfolio. The AI technologies that were developed in the context of this project produce investment signals through the components of the ASPENDYS system such as machine learning, sentiment analysis and technical analysis component. This creates a dynamic tool that can be used by investors in order to assist them in the decision-making process. The user interface is a modern web application that combines all the capabilities of a Web 3.0 application, processing information with near-human-like intelligence, through the power of AI systems, to assist end users in their investment decisions. Different data sources were used in this project to extract information regarding the assets that we studied. Specifically, social media posts from Twitter and Stocktwits as well as articles from news agencies such as BBC News, Reuters, Coin-telegraph, The Wall Street Journal and Bloomberg were used for the production of the assets' sentiment. Moreover, the Yahoo! Finance API was utilized to acquire the assets' daily values for the technical analysis and machine learning signals prediction.

During the use cases examined in this paper, we concluded that the use of the ASPENDYS platform in the investment decisions process could increase the profit of a portfolio. More specifically, in the aggressive portfolio use case, the ASPENDYS platform suggestions assisted the investor to increase their portfolio value by approximately 3%, while, in the use case of the conservative portfolio, the suggestions led to an increase of almost 5% in the portfolio value. The technologies used in the specific project implement state-of-the-art algorithms in the area of technical analysis, machine learning and sentiment analysis, a necessary feature in a modern portfolio management and model-based trading platform. In addition, the combination of all these technologies in the field of stock

market prediction positions the ASPENDYS platform in the Web 3.0 applications of the financial sector.

6. Conclusions

As mentioned in Section 2, there are several works that deal with stock prediction; most of them are based on stocks' historical values, without taking into consideration information that may affect the stock prices, such as social or political events. While referring to RQ1, the system takes into account both financial data (e.g., stocks' closing values) and textual data retrieved from either reputable news websites or social networks (i.e., Twitter and Stocktwits). For processing the financial data, different methods of technical analysis and machine learning [5,50] are utilized. The outcome of the analysis is the generation of investment signals for buying or selling stocks. Because the textual data are retrieved from various sources and in large quantities, they are firstly correlated through text analysis with the available financial symbols of the platform, and then filtered based on their reliability. Then, only the reliable data are analyzed for extracting their sentiment. By aggregating and recording the daily sentiment of each stock, time series with the sentiment of the stocks are created. By combining these time series and the price indices of the stocks, the system is able to recommend additional signals based on the general sentiment for the stocks.

Additionally, regarding RQ2, the ASPENDYS platform can serve users as a mean for managing and monitoring their investments. After being registered on the platform, the users can define a set of assets that they are interested in investing into or monitoring their course. This set of assets forms the user's portfolio. The system is capable of predicting investment signals, related to a portfolio, and alerting the end users when necessary. Finally, the platform provides, through dedicated tools, the complete analysis of the user's investment movements, presenting historical data with their investments, their portfolio profit within a certain period and the course of each asset independently.

However, there are some limitations in the proposed platform. More specifically, the fine tuning of sentiment analysis models is a difficult task because there is a lack of open source annotated datasets that contain articles and tweets in the field of economics. Therefore, as mentioned in Section 3.7.1, we have to use datasets from another sector that is a sub-optimal option. From a technical point of view, the specific implementation has not been tested in terms of scalability; however, a future plan is the deployment of the ASPENDYS platform in a business infrastructure where these kind of tests can be done. Moreover, another future plan for the proposed platform could involve the implementation and the adaptation of the algorithms and models to the cryptocurrency industry. In addition, the part of the sentiment analysis could be enhanced in the future by integrating lexicons such as the VADER (Valence Aware Dictionary for sEntiment Reasoning) lexicon [61].

Author Contributions: Conceptualization, A.D., A.K. (Athanasios Konstantinidis) and G.P.; methodology, T.-I.T., A.Z. and A.K. (Athanasios Konstantinidis); software, T.-I.T., A.Z. and T.P.; validation, M.S., A.K. (Anna Kougioumtzidou), K.T. and D.P.; formal analysis, G.P.; investigation, M.S. and A.K. (Anna Kougioumtzidou); resources, K.T. and D.P.; data curation, T.-I.T. and A.Z.; writing—original draft preparation, T.-I.T. and A.Z.; writing—review and editing, T.-I.T. and A.Z.; visualization, T.-I.T. and A.Z.; supervision, A.D.; project administration, D.T.; and funding acquisition, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the ASPENDYS project co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code:T1EDK-02264).

Data Availability Statement: All data that are not subjected to institutional restrictions are available through the links provided within the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Corbelli, R.; Vellasco, M.; Veiga, Á. Investigating Optimal Regimes for Prediction in the Stock Market. In Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, UK, 19–24 July 2020; pp. 1–8.
2. Rounaghi, M.M.; Zadeh, F.N. Investigation of market efficiency and financial stability between S&P 500 and London stock exchange: Monthly and yearly forecasting of time series stock returns using ARMA model. *Phys. A Stat. Mech. Its Appl.* **2016**, *456*, 10–21.
3. Mahmud, M.S.; Meesad, P. An innovative recurrent error-based neuro-fuzzy system with momentum for stock price prediction. *Soft Comput.* **2016**, *20*, 4173–4191. [CrossRef]
4. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef] [PubMed]
5. Borovykh, A.; Bohte, S.; Oosterlee, C.W. Conditional time series forecasting with convolutional neural networks. *arXiv* **2017**, arXiv:1703.04691.
6. Khedr, A.E.; Yaseen, N. Predicting stock market behavior using data mining technique and news sentiment analysis. *Int. J. Intell. Syst. Appl.* **2017**, *9*, 22. [CrossRef]
7. Feuerriegel, S.; Prendinger, H. News-based trading strategies. *Decis. Support Syst.* **2016**, *90*, 65–74. [CrossRef]
8. Wang, J.; Kim, J. Predicting stock price trend using MACD optimized by historical volatility. *Math. Probl. Eng.* **2018**, *2018*. [CrossRef]
9. Chong, T.T.L.; Ng, W.K. Technical analysis and the London stock exchange: Testing the MACD and RSI rules using the FT30. *Appl. Econ. Lett.* **2008**, *15*, 1111–1114. [CrossRef]
10. Liu, B. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012.
11. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
12. Esuli, A.; Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*; Citeseer: Princeton, NJ, USA, 2006; Volume 6, pp. 417–422.
13. Khoo, C.S.; Johnkhan, S.B. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *J. Inf. Sci.* **2018**, *44*, 491–511. [CrossRef]
14. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
15. Mittal, A.; Goel, A. Stock Prediction Using Twitter Sentiment Analysis. Available online: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf (accessed on 18 February 2021).
16. Ain, Q.T.; Ali, M.; Riaz, A.; Noureen, A.; Kamran, M.; Hayat, B.; Rehman, A. Sentiment analysis using deep learning techniques: A review. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 424.
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
18. Nguyen, T.H.; Shirai, K.; Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.* **2015**, *42*, 9603–9611. [CrossRef]
19. Pagolu, V.S.; Reddy, K.N.; Panda, G.; Majhi, B. Sentiment analysis of Twitter data for predicting stock market movements. In Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 3–5 October 2016; pp. 1345–1350.
20. Yu, X. Analysis of New Sentiment and Its Application to Finance. Ph.D. Thesis, Brunel University London, Uxbridge, UK, 2014.
21. Yang, X.; Liu, W.; Zhou, D.; Bian, J.; Liu, T.Y. Qlib: An AI-oriented Quantitative Investment Platform. *arXiv* **2020**, arXiv:2009.11189.
22. Hernandez-Nieves, E.; Bartolome del Canto, A.; Chamoso-Santos, P.; de la Prieta-Pintado, F.; Corchado-Rodríguez, J.M. A Machine Learning Platform for Stock Investment Recommendation Systems. In *Distributed Computing and Artificial Intelligence, 17th International Conference*; Dong, Y., Herrera-Viedma, E., Matsui, K., Omatsu, S., González Briones, A., Rodríguez González, S., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2021; pp. 303–313. [CrossRef]
23. Ren, R.; Wu, D.D.; Liu, T. Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Syst. J.* **2019**, *13*, 760–770. [CrossRef]
24. Shiller, R.J. From Efficient Markets Theory to Behavioral Finance. *J. Econ. Perspect.* **2003**, *17*, 83–104. [CrossRef]
25. Valencia, F.; Gómez-Espinosa, A.; Valdes, B. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy* **2019**, *21*, 589. [CrossRef]
26. Mohan, S.; Mullapudi, S.; Sammeta, S.; Vijayvergia, P.; Anastasiu, D.C. Stock Price Prediction Using News Sentiment Analysis. In Proceedings of the IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 4–9 April 2019; pp. 205–208. [CrossRef]
27. The Modern Web Developer's Platform. Available online: https://angular.io/ (accessed on 13 September 2020).
28. Roesslein, J. Tweepy Documentation. Available online: http://docs.tweepy.org/en/v3.5.0/ (accessed on 12 September 2020).
29. Twitter Developers. Available online: https://developer.twitter.com/en/docs/tweets/sample-realtime/guides/recovery-and-redundancy (accessed on 5 August 2020).

30. Mitra, T.; Gilbert, E. Credbank: A large-scale social media corpus with associated credibility annotations. In Proceedings of the International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015; Volume 9.
31. Buntain, C.; Golbeck, J. Automatically identifying fake news in popular twitter threads. In Proceedings of the 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, 3–5 November 2017; pp. 208–215.
32. Stocktwits API Overview. Available online: https://api.stocktwits.com/developers/docs (accessed on 7 October 2020).
33. Newsapi Documentation. Available online: https://newsapi.org/docs (accessed on 27 July 2020).
34. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. 2020. Available online: https://zenodo.org/record/4769120#.YKca8aERVPY (accessed on 10 October 2020).
35. Remote Data Access-Pandas 0.18.1 Documentation. Available online: https://pandas.pydata.org/pandas-docs/version/0.18.1/remote_data.html#remote-data-yahoo (accessed on 25 June 2020).
36. Pang, B.; Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv* **2005**, arXiv:cs/0506075.
37. Rieman, M.M.; Kennedy, A.S.; Ray, M.R. Perceived Credibility in News Depending on Author Race & Statistical Evidence. Available online: https://digitalcommons.onu.edu/student_research_colloquium/2021/papers/28 (accessed on 22 April 2020).
38. Sousa, S.; Bates, N. Factors influencing content credibility in Facebook's news feed. *Hum. Intell. Syst. Integr.* **2021**, *3*, 69–78. [CrossRef]
39. Bates, N.; Sousa, S.C. Investigating Users' Perceived Credibility of Real and Fake News Posts in Facebook's News Feed: UK Case Study. In *International Conference on Applied Human Factors and Ergonomics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 174–180.
40. Singh, R.; Choudhary, N.; Bindlish, I.; Shrivastava, M. Neural network architecture for credibility assessment of textual claims. *arXiv* **2018**, arXiv:1803.10547.
41. Segarra, S.; Eisen, M.; Ribeiro, A. Authorship attribution through function word adjacency networks. *IEEE Trans. Signal Process.* **2015**, *63*, 5464–5478. [CrossRef]
42. Markowitz, H. Portfolio Selection. *J. Financ.* **1952**, *7*, 77–91. [CrossRef]
43. Mangram, M. A Simplified Perspective of the Markowitz Portfolio Theory. *Glob. J. Bus. Res.* **2013**, *7*, 59–70.
44. Haugen, R.A. *Modern Investment Theory*, 5th ed.; Prentice Hall Finance Series: Portfolio Analysis; Prentice Hall International: Hoboken, NJ, USA, 2001.
45. Walters, C.J. The Black-Litterman Model in Detail. *SSRN Electron. J.* **2014**, 65. [CrossRef]
46. Sankaran, H.; Martin, K. Using the Black-Litterman Model: A View on Opinions. *J. Invest.* **2018**, *28*. [CrossRef]
47. Black, F.; Litterman, R. Global Portfolio Optimization. *Financ. Anal. J.* **1992**, *48*, 28–43. [CrossRef]
48. Roncalli, T.; Weisang, G. Risk Parity Portfolios with Risk Factors. *SSRN Electron. J.* **2012**, *16*. [CrossRef]
49. Maewal, A.; Bock, J. A Modified Risk Parity Method for Asset Allocation. *SSRN Electron. J.* **2018**, *3*. [CrossRef]
50. Oord, A.V.D.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
51. Krauss, C.; Do, X.A.; Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *Eur. J. Oper. Res.* **2017**, *259*, 689–702.
52. Yuan, X.; Yuan, J.; Jiang, T.; Ain, Q.U. Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market. *IEEE Access* **2020**, *8*, 22672–22685. [CrossRef]
53. Liu, J. Stock selection by using fundamental analysis and technical analysis. Available online: https://hdl.handle.net/10356/77920 (accessed on 18 May 2021).
54. Patel, K.V. A study on technical analysis with special preference to insurance sector companies with the help of MACD and RSI. *Int. J. Adv. Res. Manag. Soc. Sci.* **2019**, *8*, 71–86.
55. Farias Nazário, R.T.; e Silva, J.L.; Sobreiro, V.A.; Kimura, H. A literature review of technical analysis on stock markets. *Q. Rev. Econ. Financ.* **2017**, *66*, 115–126. [CrossRef]
56. Picasso, A.; Merello, S.; Ma, Y.; Oneto, L.; Cambria, E. Technical analysis and sentiment embeddings for market trend prediction. *Expert Syst. Appl.* **2019**, *135*, 60–70. [CrossRef]
57. Crawford, J. How to Successfully Trade Support and Resistance. Available online: https://www.learntotradeforprofit.com/support-resistance/ (accessed on 11 July 2020).
58. Investopedia. Available online: https://www.investopedia.com/ (accessed on 2 March 2020).
59. Gunasekarage, A.; Power, D.M. The profitability of moving average trading rules in South Asian stock markets. *Emerg. Mark. Rev.* **2001**, *2*, 17–33. [CrossRef]
60. Spyromitros-Xioufis, E.; Tsoumakas, G.; Groves, W.; Vlahavas, I. Multi-target regression via input space expansion: Treating targets as inputs. *Mach. Learn.* **2016**, *104*, 55–98. [CrossRef]
61. Sohangir, S.; Petty, N.; Wang, D. Financial sentiment lexicon analysis. In Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 286–289.

future internet

MDPI

Article

From Rigidity to Exuberance: Evolution of News on Online Newspaper Homepages

Simón Peña-Fernández *, Miguel Ángel Casado-del-Río and Daniel García-González

Faculty of Social and Communication Sciences, University of the Basque Country (UPV/EHU), Barrio Sarriena, 48940 Leioa, Spain; miguelangel.casado@ehu.eus (M.Á.C.-d.-R.); daniel.garcia@ehu.eus (D.G.-G.)
* Correspondence: simon.pena@ehu.eus

Abstract: Since their emergence in the mid-90s, online media have evolved from simple digital editions that merely served to dump content from print newspapers, to sophisticated multi-format products with multimedia and interactive features. In order to discover their visual evolution, this article conducts a longitudinal study of the design of online media by analyzing the front pages of five general-information Spanish newspapers (elpais.com, elmundo.es, abc.es, lavanguardia.com, and elperiodico.com) over the past 25 years (1996–2020). Moreover, some of their current features are listed. To this end, six in-depth interviews were conducted with managers of different online media outlets. The results indicate that the media analysed have evolved from a static, rigid format, to a dynamic, mobile, and multi-format model. Regarding the language used, along with increased multimedia and interactive possibilities, Spanish online media currently display a balance between text and images on their front pages. Lastly, audience information consumption habits, largely superficial and sporadic, and the increasing technification and speed of production processes, means that news media have lost in terms of the design part of the individual personality they had in their print editions. However, they maintain their index-type front pages as one of their most characteristic elements, which are very vertical and highly saturated.

Keywords: online media; design; front pages; digital editions; journalism; web 3.0; visualisation; multimedia; interaction

Citation: Peña-Fernández, S.; Casado-del-Río, M.Á.; García-González, D. From Rigidity to Exuberance: Evolution of News on Online Newspaper Homepages. *Future Internet* **2021**, *13*, 150. https://doi.org/10.3390/fi13060150

Academic Editors: Andreas Veglis and Charalampos Dimoulas

Received: 28 April 2021
Accepted: 7 June 2021
Published: 9 June 2021

1. Introduction

Throughout history, the design of print newspapers has maintained a long tradition of conventions, rules, and standards for the display and hierarchization of information with which readers were familiarized for decades [1–3].

However, since the emergence of the Internet and the first online media in the mid-90s, the media has constantly had to confront challenges such as permanently updating information, multiplied content and formats, and the appearance of new formats and screens for reading [4]. During the 25 years since then, their evolution has been characterized by the wide range of multimedia and multi-format possibilities [5] and their ongoing mutation, encouraged by constant technological change [6]. "The main feature of this format is that it possibly does not have a final model. There shall be no newspaper of the future, but rather a newspaper of a determined present, conditioned by specific technologies in constant movement and cultural and professional concepts. There is no final destination, and no model that stabilizes with success. The new paradigm is constant evolution. The final format may be constant change" [7].

Despite the broad interest sparked amongst researchers in the online media field of study, design is one of the scopes that has received the least attention [8], although there are some monographs that specifically address this issue [9–14] and different specific studies on the design of digital newspapers [15,16]. Research has also been carried out on the user experience and engagement of online media, and the way news is read on different devices [17–21].

In this evolutive context, the goal of this article is to analyze the main visual transformations that online media have undergone in their first 25 years of history. Specific research questions are to discover the specific characteristics of current online news consumption (RQ1), to determine their textual and/or visual nature (RQ2), and to characterize their main features from a journalistic-design perspective (RQ3).

2. Evolution of the Design of Online Media

2.1. From Dumping to Webpage (1996–2000)

The appearance of the first digital editions on the Internet in the mid-1990s was characterized by simply dumping print-publication content to the new format [22], without generating new content or developing new design criteria and hierarchization. The lack of specific principles means that this stage can be defined as "proto design" stage for digital newspapers [23]. Generally, these first editions fundamentally met the professional and corporate imperative of occupying a space in the new environment, without excessive concern for how this objective was achieved [24].

During these initial years, far from having a standard model [25], the media experimented, to greater or lesser success, with different kinds of resources (frames, buttons, etc.) to present news, many of which rapidly fell into disuse, during turbulent beginnings when newspapers were quickly redesigned [26] (Figure 1).

Figure 1. Print-publication dumping in early abc.es online edition (27 March 1997). Compiled by the authors.

Beginning in 1998, the media began generating complementary content (and not only exclusively journalist content) to attempt to attract an audience that was gradually peering into the new news window. These webpages, which had found new competitors in the webpages created by telecommunications operators, aimed to develop a channel for commerce and advertising in parallel, combining information, entertainment, and services [7].

In 1998, the arrival of dynamic HTML facilitated organization of contents on the page, and the architecture began to display side or top navigation menus, and also long vertical areas where contents were organized [23]. Moreover, the availability of cascade-style sheets (CSS) helped with separating the structure of the documents from their formal presentation [26].

2.2. From Digital Editions to Online Media (2001–2007)

Although the beginnings were faltering, this lack of definition did not last long, and changes were quick to come. In a bit more than 5 years after their creation, all the main newspapers had pages where the content specifically created for the web was pushing over information from the print editions and beginning to steal the spotlight. Online media thus began to emancipate from the analogical media from whence they had come and that had inspired them, and began developing their own narrative and visual, specifically digital, characteristics. One could no longer speak of simple adaptations or digital editions of pre-existing media, but rather of online media with their own, independent personality.

While during these years technical limitations made it difficult for images to find their space on media webpages, more varied graphic resources began to appear, with the incipient presence of videos [27]. In the page architecture, the most customary structure during this period was called "Inverted L" or "Trident," splitting the page into three areas: the masthead and general navigation menu were placed on top, while the detailed menu was to the left, and the news occupied the main part of the screen, organized in vertical fashion [23] (Figure 2).

Figure 2. "Trident" structure in elperiodico.com and elmundo.es (August 2002). Compiled by the authors.

2.3. Mobile Devices Come into Play (2007–2014)

After a decade of consuming online media almost exclusively through personal computer monitors, the launch of the iPhone in 2007 and the emergence of the first tablets were the most emblematic milestones in a new way of consuming information that was quick to spread: mobile phones [28]. Although the first mobile interfaces had appeared in the late 90 s and there were already alternative information formats like the PDA (Personal Digital Assistant) and communication technologies like SMS (Short Message Service) and MMS (Multimedia Message Service), their success up until that point had been meagre [29,30]. The reduced use of these devices, their small screen size, and the low capacity of mobile data networks had limited these initiatives to the creation of simple content, complementary in nature to the traditional media discourse [31] (Figure 3).

Figure 3. elmundo.es app (24 July 2012) (accessed on 17 April 2021). Compiled by the authors.

However, the spread of the 3G (2000) and 4G (2010) data networks and the popularization of smartphones contributed to the emergence of exclusive downloadable applications and the development of exclusive narratives and formats that coexisted with content originally created for other formats. In addition to its ubiquitous nature, this fourth screen [32] was multimedia in nature and had interactive and convergent potential, providing for double or non-exclusive news consumption.

2.4. Adaptable Web Design and Apps (2014–Current)

Since then, the emergence and later success of mobile devices as a format for news consumption has led news media to prioritize rapid, versatile publication of contents in all kinds of formats and media. The number of platforms from whence information can be consumed has multiplied, and all of them (desktop computer, laptop, tablet, mobile phone, watches, etc.) need their own specific format.

The first response to this need was adaptive web design [33], with which several design versions of the page are created in different formats adapted to the most common screen widths (for example, 320 px, 480 px, 760 px, etc.). The website detects the screen size of each device that connects and offers the option that best fits that size with previously programmed style sheets [34]. This is all in a complex set of languages into which the webpages are structured in HTML5, giving them a style with the CSS, and adding functionalities with Javascript [5].

Adaptive design, however, bears certain limitations. Since it is based on rigid templates, it forces one to make different designs for each one of the most customary screen widths. For this reason, regardless of the size of the screen making the connection, the website can only offer one of the pre-established designs.

The adaptive design evolved, giving birth to responsive web design [34,35], which allows one to fluidly adjust the content of the website to any device. To this end, instead of using predefined templates, grids that automatically adjust in number and percentage size to the window used to view the content are used, which in practice means a practically unlimited number of viewing options.

These adaptable web design modalities, whether adaptive or responsive, provide for content that adjusts to very different screens (mobiles, tablets, monitors) from whence news is accessed, and are the key to the success of mobile and multi-platform news consumption. On the other hand, from the perspective of media design, adaptable web design has limited options for journalists to control page architecture and how users view the information, preferring accessibility and ease in news consumption. On the other hand, content distribution can substantially vary from one browser to another, which creates completely different user experiences with one same product.

In parallel fashion, in this last stage, mobile apps also appeared for the consumption of online news [36]. More than anything, they provide for more complete multimedia integration of content [37]. Apps are the first real alternative to the almost exclusive consumption of news through web browsers.

For the media, apps are formats that garner high loyalty, since they are a source of monomedia consumption; however, for the time being, they have not yet managed to take over a significant portion of news consumption. Their complex development, their exclusive nature (with a specific app for each media outlet), constant changes in formats, and the consequent need to update apps, as well as the impossibility of showing all formats, has limited their implementation until now [4].

3. Materials and Methods

To analyze the visual evolution of Spanish online media (Figure 4), the analysis of five general-information news media that share a conjoint trajectory throughout the entire analyzed period were used as a base: elpais.com, elmundo.es, abc.es, lavanguardia.com and elperiodico.com (accessed on 17 April 2021).

Figure 4. Evolution of El País homepage. Compiled by the authors.

In total, 125 front pages were analyzed, one per year from each one of the selected media outlets between 1996 and 2020. The front pages were viewed with the digital Internet Archive [38], which stores original screenshots of the webpages, given that the media's digital newspaper archives have adapted old content to the current design [16].

To analyze the front pages, we considered existing methodological proposals to analyze digital media [27,39,40], based on which a file was drawn up that included the number of news pieces and images; the total page surface area, with special attention to the verticality; and the graphic surface area as variables for the longitudinal study of the design. Journalistic content (news and opinion) were accounted for, and commercial contents were not considered. Moreover, the study is focused on studying front-page elements and does not consider other issues such as page architecture and user experience.

To describe the media consumption habits, the data from the reports published annually by the Association for Media Research (AIMC: Madrid, Spain) [41] have been used. For its 2020 report, 29,097 interviews were conducted with the population residing in Spain over 14 years of age. The data on the number of visits to the web pages and their origin have been obtained from Similarweb (2021) [42].

In parallel, six in-depth interviews were conducted with news media managers to learn the current design features in Spanish online media. These include traditional media (elpais.com, elmundo.es), native media (diario.es, elconfidencial.com, elespanol.com), and audio-visual media (rtve.es) (accessed on 17 April 2021).

4. Results

4.1. Web, Mobile, and Multi-Format Consumption

While Spanish online media outlets came to be 25 years ago as complementary digital versions of print media that simply dumped their content and were mainly consumed

from desktop computers, the emergence of smartphones and tablets, as well as increased telephone network data capacity, radically modified these habits.

Thus, online media are now leading formats in the consumption of news, much more so than traditional media (Figure 5). According to the AIMC, in 2019, 51.1% of newspaper audiences access news content exclusively through the internet, as opposed to 31.5% who consume exclusively on paper, and 17.4% who use both formats. This inversion in format preference has not reduced total consumption of news; rather, to the contrary, it has led to total newspaper audiences rising from 36.5% of the population in 2000 to 42.5% in 2019 [41].

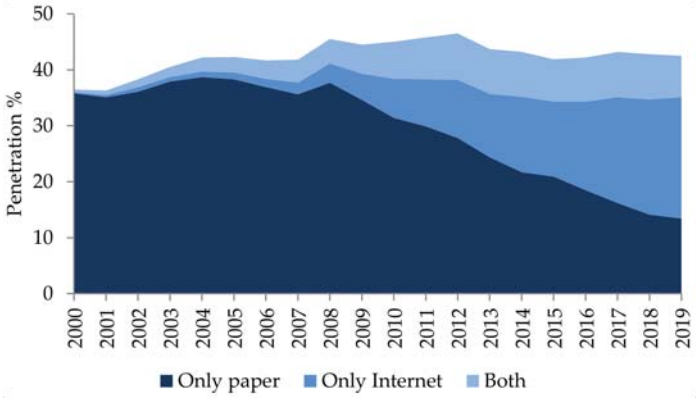


Figure 5. Evolution of newspaper audiences according to format (Penetration %). Compiled by the authors on the basis AIMC (2020) [41] data.

Secondly, while consumption is now mainly online, the desktop computer has lost main device status for consulting news online. In the past 5 years, audience evolution by device [41] shows that while consumption from personal computer has slightly lowered (from 13.6% to 11.1%), on smartphones it has doubled, reaching 23.5% (Figure 6). As such, the mobile phone is now very much the main device for consulting online news pieces (58.9%), ahead of computers (28%), tablets (12.8%), and other devices (0.3%).

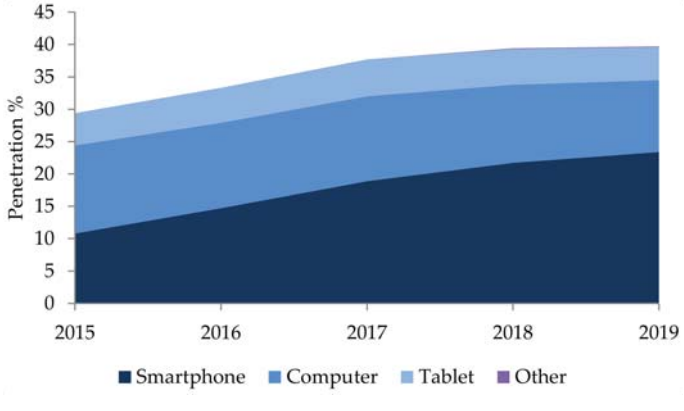


Figure 6. Evolution of newspaper audiences by format device (Penetration %). Compiled by the authors on the basis AIMC (2020) [41] data.

News media managers corroborate this about-face in news consumption habits, which, in addition to other aspects, entailed a change in the time distribution of audiences, with the emergence of a great peak in information consumption from mobile devices first thing

at night, which now joins first thing in the morning (in this case, with most consumption from the computer) and at noon (Interviews I1, I2, I3, I4, I5, I6). The transformation of newsrooms from a paper model, with a closing time to prepare a morning edition, to a 24-h news flow, also forced them to change their work routines.

Thirdly, in a news ecosystem where almost everything has changed, one element remains unalterable: among the readers of the newspapers, online news is still mainly consumed through web browsers (76.7%), much more than apps (19.8%), and other information viewers (3.5%) that have not taken off as news consultation devices (Figure 7).

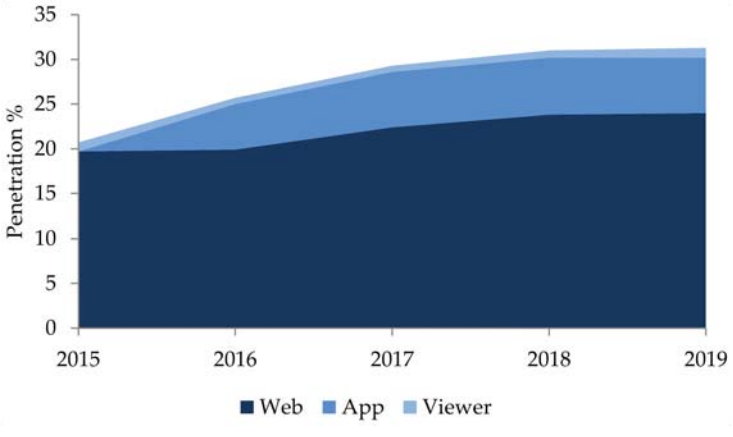


Figure 7. Evolution of newspaper audiences by format device (Penetration %). Compiled by the authors on the basis AIMC (2020) [41] data.

4.2. Balance between Text and Images

The history of online news has also been a story of evolution, from an almost exclusively written model to another model where text and images have a balanced, complementary presence. In this section, technological evolution was a powerful conditioning factor on visual evolution, because while the limited internet bandwidth capacity initially fostered heavy text content, today, the balance between text and images reigns, with a model where practically each front-page news piece is accompanied by its own image. Thus, the ratio that relates the number of images to the number of news currently displays an almost complete balance (1:1.2) (Figure 8).

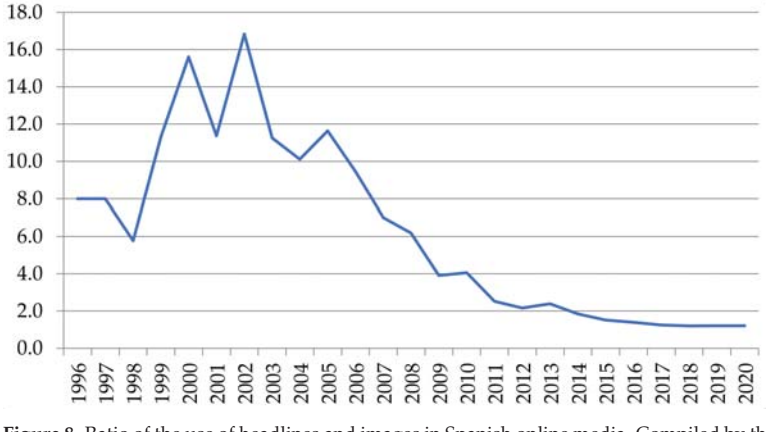


Figure 8. Ratio of the use of headlines and images in Spanish online media. Compiled by the authors.

Amongst texts, headlines have taken on almost exclusive protagonism in designs. Although in the first years imitations of print references led to the inclusion of subheadlines and small text lead-ins on front pages, barring some cases with the top news on the page, practically all front-page news pieces on online media simply have a headline and an image, without any other text accompaniment.

Additionally, the protagonism of images has radically increased. While the limitations of existing networks and the visual legacy of print media made photo news practically the only highlighted visual element on front pages during the first 10 years, in the past 5 years, the analyzed media outlets' front pages displayed practically 100 images associated with news content.

This radical transformation in the distribution of content cannot be understood without visual protosensitivity in the media, given that, since their beginnings, they linked visual content on their front pages, although this content could not be viewed in integrated fashion from the front pages themselves.

4.3. Saturation and Verticality

Adaptable web design and mobile and multi-format consumption led to online media losing part of the visual singularity that characterized print media, preferring accessibility and adaptability of content. The increasing technification of production processes (with the need to combine different programming languages) took journalists away from online media design tasks, which are now in the hands of web programmers.

In an enormously complex setting of devices and formats, online media prioritize making professionals' work easier in creating multi-platform content by implementing content management systems (CMS) to simplify content creation and publication tasks to the maximum. While self-publishing programs such as QuarkXPress and Adobe InDesign grant newsrooms control, practically without technical mediation in the process of preparing print newspapers, with the incorporation of the web, the number of languages multiplies, which requires specialized staff with great knowledge of programming, and brings the work of journalists to platforms with limited design options.

However, despite losing part of this visual singularity that characterized print media, online media have incorporated a very characteristic visual trait that sets them apart from other webpages: high saturation of news pieces on extremely vertical homepages.

During their first 10 years of history, homepages scarcely increased in size, and news pieces were almost exclusively text. However, the progressive incorporation of more news pieces on front pages, as well as their accompanying images, fostered an exponential increase in the size of these front pages, multiplying their original size by up to 10 times (Figure 9).

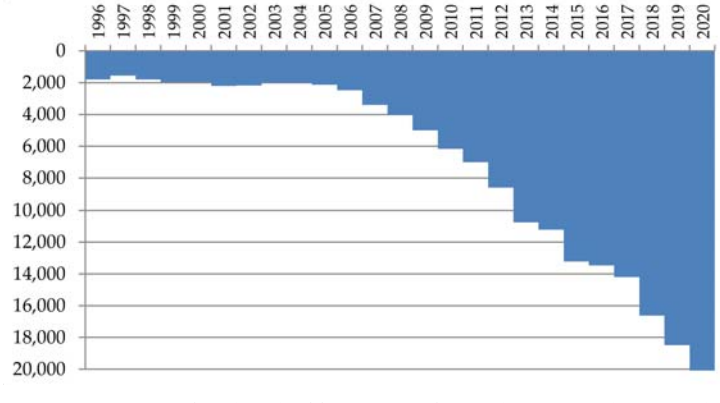


Figure 9. Evolution of the length of front pages of Spanish online media (in pixels). Compiled by the authors.

Despite attempts to seek out a certain balance in the horizontal layout of news content on the page, today, online media are characterized by the extreme verticality of their index-type front pages. Thus, during the first 10 years of history, online media showed a more contained attitude, and their covers included less than 50 news items a day, but since 2010 the trend of publishing a much larger number of content has increased significantly, displaying on average over 100 news pieces and images (in 2020, 134.6 and 111.8, respectively). This trend is seen equally in all the media, although in the case of ABC, the number currently reaches almost 200 (Figure 10).

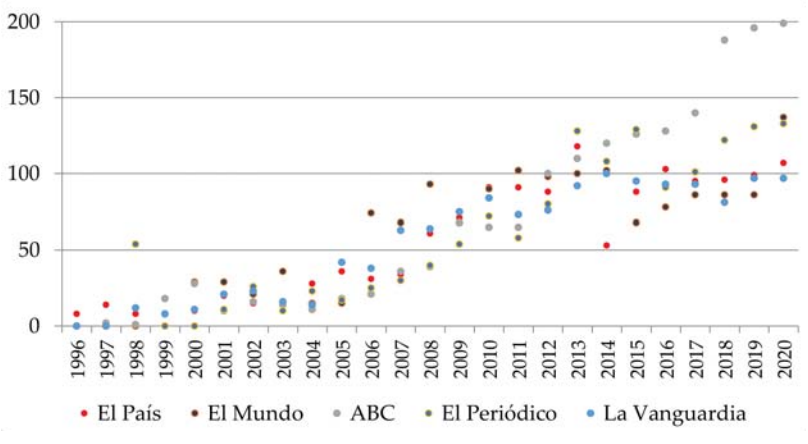


Figure 10. Number of news pieces on the front page of Spanish online media. Compiled by the authors.4.4. Front pages and SEO as access windows.

Pages to access online media have therefore evolved from a design with a selective display case of content that emulated the front pages of print newspapers, to another, laid out as an exhaustive meta-front page for all sections of the newspaper.

The exuberant supply of content contrasts with readers' consumption habits. For the time being, general data indicate that reading news on online media continues to be brief and superficial. If we consider three basic indicators (Table 1) (average visit time, average viewed pages, and bounce rate), we see epidemic consumption of the news pieces supplied. Thus, for the five media outlets analyzed, the average visit duration is 6 min and 16 s, the average pages viewed per visit is 2.9, and the "bounce rate" (meaning, the percentage of visitors who invest less than 30 s in the website before going to a different one) is 54.8% (Table 2). Higher consumption through mobile phones has worsened this trend, given that they reduce the time spent on reading in comparison with desktop computers.

Table 1. Media circulation and visits. Compiled by the authors on the basis of OJD (2020) [43] and Similarweb (2021) [42] data.

	Circulation (Paper)	Visits per Month (Web)
elpais.com	72.471	137.95 M
elmundo.es	45.111	113.05 M
abc.es	53.436	81.10 M
lavanguardia.com	73.296	78.81 M
elperiodico.com	33.506	27.67 M

Table 2. Online media access channels. Compiled by the authors on the basis of Similarweb (2021) [42] data.

	Average Time	Pages per Visit	Bounce Rate
elpais.com	6:27	2.32	59.44%
elmundo.es	6:17	3.27	52.26%
abc.es	5:36	2.85	50.32%
lavanguardia.com	5:50	3.28	59.11%
elperiodico.com	7:12	2.82	52.80%

These consumption habits have at least two implications on design. On one hand, we must also consider people accessing specific news pieces without going through the front page in this average; for example, through links shared on social media. On the other, they explain that online media front pages decide to supply the maximum number of news pieces possible on their front page, so that the visitor who is going to devote a limited amount of time to reading the media outlet can not only do a quick scan of current news but can also quickly find the pieces they wish to read in depth.

Access channels to online media also explain these consumption habits (Figure 11). In all analyzed media outlets, 38.9% of visits directly access the webpage, writing the name in the address bar, while access through search engines is 49.3% (largely by searching for the name of the outlet), and social media is at 5.6%.

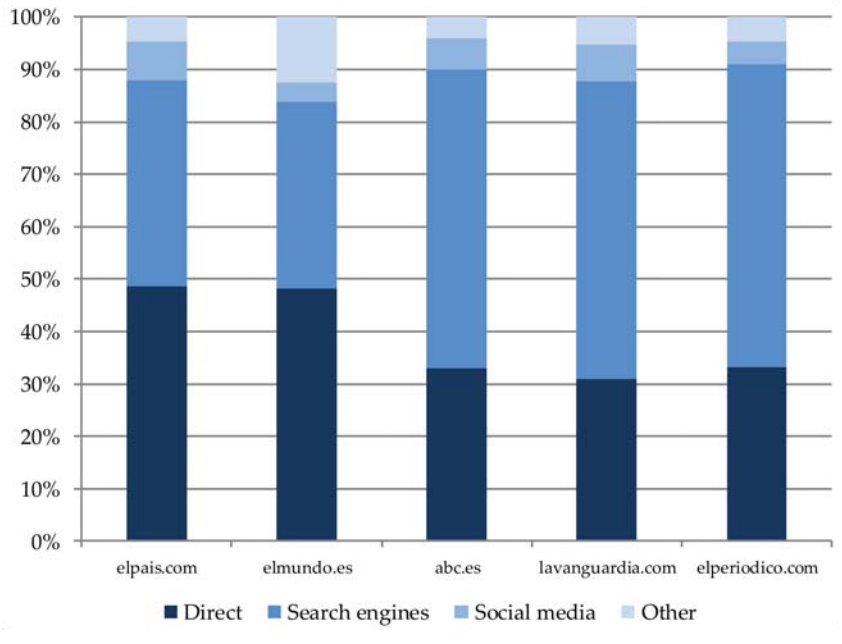


Figure 11. Online media access channels. Compiled by the authors on the basis Similarweb [42] data.

As such, the media's front pages have lost their status as the exclusive window to access the newspapers. While mobile consumption has encouraged more superficial searches for content that foster the creation of highly saturated index-type front pages, new channels to access contents (search engines and social media) have made journalistic SEO a new window to access news.

Monitoring traffic generated by each one of the news pieces is capitally important for online newsrooms (Interviews I1, I3, I5, I6), although news media managers indicate that

the main objective is knowledge of the most loyal users, and not sporadic visitors that may congregate around news pieces. In turn, this policy is aligned with subscription payment policies, which almost all Spanish online media now have.

Regarding hierarchization criteria for news pieces, current events continue to be at the top of front pages, while content with a less journalistic, more lightweight profile (technology, services), are concentrated on the lower half of pages. In the middle, after current events, are highlighted recommended articles (features, interviews) and opinion texts. Hierarchization criteria for news pieces continue to be purely professional, and the media refuse to use automatic configurations to organize their content (Interviews I1, I2, I3, I4, I5).

5. Discussion and Conclusions

In the past 25 years, design of online media has evolved from simple complementary editions of the analogical media that housed a mere transfer of content, to complex online media distributed in multi-format mode by an adaptable web design.

In terms of design, growing technification of the process of displaying news pieces, and the multi-format nature of their distribution, have sacrificed part of the visual singularity that characterized the press, and left media design in the hands of staff who often lack journalistic or design training [44], while journalists work with content management systems (CMS) [45]. However, we must not forget that the design of journalist news has always entailed different disciplines working together for a purpose [46,47].

While traditional press design bore front pages that were enormously synthetic and highly hierarchized, governed by very stable conventions, today's online media are governed by generic web design patterns, such as usability [48], information architecture [49], interaction [50–52], and user experience [5].

Despite this, news sites can be considered as a specific type of website that shows some peculiarities. The use of the web has provided the media with specific characteristics such as rapid content updating, multimedia or interaction, which have transformed, over the years, the way in which information is presented. Thus, online newspaper homepages have evolved from a rigid model to a dynamic one, with its own visual identity. In this regard, the most distinctive element of online media design is the presence of highly vertical and saturated index-type front pages, which continue to operate as display cases for an exuberant supply of news pieces [16], which forces constant displacement in order to view the profuse supply of content [53].

In terms of content, front pages have evolved from an almost exclusively text model to another model characterized by balance between text and images, which indicates a growing visual trend [15] and underlines the fundamental role that images play in online journalism, ahead of other multimedia formats [54]. Thus, the convergence of content that pivoted around the use of text as a nuclear element that characterizes news media webpages has declined [31].

Moreover, the design of online media has been conditioned by changes in audiences' reading habits, going from customary consumers of one sole format to broadening their customary news diet to a greater number of media outlets, which has fostered an exponential increase in sporadic and superficial consumption of news [55]. This has reconfigured front pages, which have gone from being selective display cases to meta front pages with sections and indices for news pieces.

Additionally, front pages have lost their status as the exclusive access point to news pieces since online media now conjointly draw in over half of their readers through search engines and social media. For this reason, SEO techniques rival traditional criteria for relevance and hierarchization in the most relevant sources to access the media [56]. In visual terms, we can state that while journalistic design was focused on the design of newspapers, in online media, the design of journalistic information takes care of information as a new unit of value [57].

In the immediate future, the beginning of a new visual transformation of online media will be influenced by factors such as the new forms of non-linear storytelling, sharing and authoring [58] or the use of augmented reality [59], while the development of subscription models, with more loyal audiences that directly access the media, may be the main future challenge in this sphere [60].

Author Contributions: Conceptualization, S.P.-F.; methodology, S.P.-F.; formal analysis, S.P.-F. and M.Á.C.-d.-R.; investigation, S.P.-F. and D.G.-G.; resources, D.G.-G.; data curation, S.P.-F. and M.Á.C.-d.-R.; writing—original draft preparation, S.P.-F. and D.G.-G.; writing—review and editing, S.P.-F.; visualization, D.G.-G.; supervision, S.P.-F.; funding acquisition, S.P.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National R+D+i Plan of the Spanish Ministry of Science, Innovation, and Universities, and by the European Regional Development Fund (ERDF), grant number RTI2018-095775-B-C41.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.aimc.es/a1mc-c0nt3nt/uploads/2020/01/marco2020.pdf (accessed on 17 April 2021) Other data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reasons.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Evans, H. *Diseño y Compaginación de la Prensa Diaria*; Gustavo Gili: Ciudad de México, México, 1995.
2. Zorrilla-Ruiz, J. *Introducción al Diseño Periodístico*; EUNSA: Pamplona, Spain, 1997.
3. González-Díez, L.; Pérez-Cuadrado, P. *Principios Básicos Sobre Diseño Periodístico*; Editorial Universitas: Madrid, Spain, 2001.
4. Peña-Fernández, S.; Lazkano-Arrillaga, I.; García-González, D. La transición digital de los diarios europeos: Nuevos productos y nuevas audiencias. *Comunicar* **2016**, *46*, 27–36. [CrossRef]
5. Yunquera-Nieto, J. Tabletas y smartphones. El diseño editorial obligado a adaptarse a los nuevos soportes informativos digitales. *adComunica* **2015**, *9*, 133–155. [CrossRef]
6. Serrano, A. Del diseño gráfico y audiovisual al diseño de interacción: Un estudio sobre los nodos iniciales de los cibermedios. In *Nuevos Medios, Nueva Comunicación. Libro de Actas del II Congreso Internacional de Comunicación 3.0*; Ortega, F., Cardeñosa, L., Eds.; Universidad de Salamanca: Salamanca, Spain, 2011; pp. 159–176.
7. Gago, M. Flash 2.0. Tecnología y cibermedios en la nueva web social. In *Diseño Periodístico en Internet*; Larrondo, A., Serrano, A., Eds.; UPV/EHU: Bilbao, Spain, 2007; pp. 103–128.
8. Salaverría, R. Ideas para renovar la investigación sobre medios digitales. *Prof. Inf.* **2015**, *24*, 223–226. [CrossRef]
9. Canga-Larequi, J.A.; Coca-García, C.; Martínez-Rivera, E.; Cantalapiedra-González, M.J.; Martínez-Odriozola, L. *Diarios Digitales. Apuntes Sobre un Nuevo Medio*; UPV/EHU: Bilbao, Spain, 2000.
10. Armentia, J.I.; Elexgaray, I.; Pérez, J.C. *Diseño y Periodismo Electrónico*; UPV/EHU: Bilbao, Spain, 1999.
11. Larrondo-Ureta, A.; Serrano-Tellería, A. *Diseño Periodístico en Internet*; UPV/EHU: Bilbao, Spain, 2007.
12. Serrano, A. Diseño de Nodos Iniciales en Cibermedios: Un Estudio Comparativo. Ph.D. Thesis, University of the Basque Country, Leioa, Spain, 2010.
13. López, R. *Diseño de Periódicos y Revistas en la era Digital*; Fragua: Madrid, Spain, 2013.
14. Rodríguez-Barbero, M. Análisis y Estudio de la Arquitectura de la Información en los Cibermedios Extremeños. Ph.D. Thesis, Universidad de Extremadura, Badajoz, Spain, 2015.
15. Cabrera, M.A. El diseño de la prensa digital española en el contexto de la convergencia tecnológica. La identidad visual del ciberperiodismo. *Rev. Latina de Com. Soc.* **2009**, *64*. [CrossRef]
16. Cea-Esteruelas, N. Estudio evolutivo del diseño periodístico en Internet: La edición digital de El País. *ZER* **2014**, *37*, 137–155. Available online: https://ojs.ehu.eus/index.php/Zer/article/view/13532 (accessed on 17 April 2021).
17. Barnhurst, K.G. The Form of Online News in the Mainstream US Press, 2001–2010. *J. Stud.* **2012**, *13*, 791–800. [CrossRef]
18. Aranyi, G.; van Schaik, P. Testing a model of user-experience with news websites. *J. Assoc. Inf. Sci. Tech.* **2016**, *67*, 1555–1575. [CrossRef]
19. He, C.; Chen, N.; Zhou, M.; Li, H.; Chen, K.; Guan, D. Improving Mobile News Reading Experience for Chinese Users: An User Interview and Eye Tracking Study. *Int. Conf. Hum. Comput. Interact.* **2019**, *11585*, 395–412. [CrossRef]
20. Kim, H.K.; Jeon, H.; Choi, J. How does the design element of a news website influence user experience? *ICIC Ex. Let.* **2020**, *14*, 265–271. [CrossRef]
21. Lu, Y.; Wang, X.; Ma, Y. Comparing user experience in a news website across three devices: Iphone, ipad, and desktop. *Proc. ASIST Annu. Meet.* **2013**, *50*. [CrossRef]

22. Díaz-Noci, J.; Meso-Ayerdi, K. Tipología de los medios de comunicación en Internet. Génesis y desarrollo de un nuevo paradigma comunicativo. El caso vasco. In *XIV Congreso de Estudios Vascos "Informazioaren Gizartea = Sociedad de la Información = Société de l'Information"*; Maxwell, C., Ed.; Eusko Ikaskuntza: San Sebastian, Spain, 1998; pp. 77–83.
23. Armentia, J.I. La lenta evolución del diseño periodístico en la Red. In *Diseño Periodístico en Internet*; Larrondo, A., Serrano, A., Eds.; UPV/EHU: Bilbao, Spain, 2007; pp. 31–60.
24. Hassan, Y. Experiencia del usuario y medios de comunicación en Internet. In *Diseño Periodístico en Internet*; Larrondo, A., Serrano, A., Eds.; UPV/EHU: Bilbao, Spain, 2007; pp. 129–146.
25. Canga-Larequi, J.A. Periodismo en la Red. Diseño periodístico y ediciones digitales. *Telos* **2005**, *63*, 71–76.
26. Salaverría, R.; Sancho, F. Del papel a la Web. Evolución y claves del diseño periodístico en internet. In *Diseño Periodístico en Internet*; Larrondo, A., Serrano, A., Eds.; UPV/EHU: Bilbao, Spain, 2007; pp. 207–239.
27. Caminos, J.M.; Marín, F.; Armentia, J.I. Novedades en el diseño de la prensa digital española (2000–2008). *Palabra Clave* **2008**, *11*, 253–269.
28. Aguado, J.M.; Martínez, I.J. La comunicación móvil en el ecosistema informativo: De las alertas SMS al Mobile 2.0. *Trípodos* **2008**, *24*, 107–118.
29. Westlund, O. Mobile News. *Digit. J.* **2013**, *1*, 6–26. [CrossRef]
30. Westlund, O. The Production and Consumption of News in an Age of Mobile Media. In *The Routledge Companion to Mobile Media*; Goggin, G., Hjorth, L., Eds.; Routledge: New York, NY, USA, 2014; pp. 135–145.
31. Canavilhas, J. Contenidos informativos para móviles: Estudio de aplicaciones para iPhone. *Textual Vis. Media* **2009**, *2*, 61–80.
32. Aguado, J.M.; Martínez, I.J. Construyendo la cuarta pantalla. Percepciones de los actores productivos del sector de las comunicaciones móviles. *Telos* **2009**, *83*, 62–71.
33. Perkowitz, M.; Etzioni, O. Adaptative Web Sites: An IA challenge. In Proceedings of the 15th International Joint Conference on Artifical intelligence IJCAI-97, Nagoya, Japan, 23–29 August 1997; pp. 16–21.
34. Gustafson, A. *Adaptive Web Design: Crafting Rich Experiences with Progressive Enhancement*; New Riders Publishing: San Francisco, CA, USA, 2015.
35. Frain, B. *Responsive Web Design with HTML5 and CSS3*; Packt Publishing: Birmingham, UK, 2012.
36. Ortega, F.; González-Ispierto, B.; Pérez-Peláez, M.E. Audiencias en revolución, usos y consumos de las aplicaciones de los medios de comunicación en tabletas y teléfonos inteligentes. *Latina* **2015**, *70*, 627–651. Available online: https://www.doi.org/10.4185/RLCS-2015-1063 (accessed on 17 April 2021).
37. Costa-Sánchez, C.; Rodríguez-Vázquez, A.I.; López-García, X. Medios de comunicación móviles. Potencialidades de las aplicaciones para Smartphone de los medios de comunicación españoles de mayor audiencia. *Prism. Soc.* **2015**, *15*, 387–414.
38. Internet Archive. Wayback Machine. 2021. Available online: https://archive.org/ (accessed on 17 April 2021).
39. Cabrera, M.A.; Palomo, B. Metodologías de investigación en diseño periodístico. In *Ciberperiodismo. Métodos de Investigación*; Díaz-Noci, J., Palacios, M., Eds.; UPV/EHU: Leioa, Spain, 2008; pp. 52–62.
40. Rodríguez-Martínez, R.; Codina, L.; Pedraza-Jiménez, R. Cibermedios y web 2.0: Modelo de análisis y resultados de aplicación. *Prof. Inf.* **2010**, *19*, 35–44. [CrossRef]
41. AIMC (Asociación para la Investigación de Medios de Comunicación). *Marco General de los Medios en España*; AIMC: Madrid, Spain, 2020.
42. Similarweb. Top Websites Ranking. 2020. Available online: https://www.similarweb.com/top-websites/ (accessed on 17 April 2021).
43. OJD (Oficina de Justificación de la Difusión). *Auditoria de Medios Impresos*; OJD: Madrid, Spain, 2020. Available online: https://www.ojd.es/portfolio/auditoria-de-medios-impresos/ (accessed on 17 April 2021).
44. Machin, D.; Polzer, L. *Visual Journalism*; Macmillan International Higher Education: New York, NY, USA, 2015.
45. López, J.; Torregrosa, J.F. Rutinas productivas de los diarios digitales españoles: Caracterización y desarrollo en la dinámica de la convergencia. *Ámbitos* **2013**, *22*, 156. Available online: http://institucional.us.es/ambitos/?p=156 (accessed on 17 April 2021).
46. Pereira, X. Arquitectura de la información. Ingeniería del periodismo. In *Diseño Periodístico en Internet*; Larrondo, A., Serrano, A., Eds.; UPV/EHU: Bilbao, Spain, 2007; pp. 193–206.
47. González-Díez, L.; Puebla-Martínez, B.; Pérez-Cuadrado, P. De la maquetación a la narrativa transmedia. Una revisión del concepto de 'diseño de la información periodística'. *Palabra Clave* **2018**, *21*, 445–468. [CrossRef]
48. Nielsen, J. *Designing Web Usability: The Practice of Simplicity*; New Riders Publishing: Thousand Oaks, CA, USA, 1999.
49. Rosenfeld, L.; Morville, P. *Information Architecture for the World Wide Web*; O'Reilly: Sebastopol, CA, USA, 1999.
50. Linares, J.; Codina, L.; Vàllez, M.; Rodríguez-Martínez, R. *Interactividad, Buscabilidad y Visibilidad en Cibermedios: Sistema de Análisis y Resultados*; Serie DigiDoc: Barcelona, Spain, 2016.
51. Freixa, P.; Pérez-Montoro, M.; Codina, L. Interacción y visualización de datos en el periodismo estructurado. *Prof. Inf.* **2017**, *26*, 1076–1090. [CrossRef]
52. Codina, L.; Gonzalo-Penela, C.; Pedraza-Jiménez, R.; Rovira, C. *Posicionamiento Web y Medios de Comunicación Ciclo de Vida de una Campaña y Factores SEO*; Serie DigiDoc: Barcelona, Spain, 2017. [CrossRef]
53. Peña-Fernández, S.; Pérez-Dasilva, J.A.; Genaut-Arratibel, A. Tendencias en el diseño de los diarios vascos y navarros en Internet. *Mediatika* **2010**, *12*, 105–137.

54. Guallar, J.; Rovira, C.; Ruiz, S. Multimedialidad en la prensa digital. Elementos multimedia y sistemas de recuperación en los principales diarios digitales españoles. *Prof. Inf.* **2010**, *19*, 620–629. [CrossRef]
55. Milosevic, M.; Chisholm, J.; Kilman, L.; Henriksson, T. *World Press Trends 2014*; WAN-IFRA: Paris, France, 2014.
56. Iglesias-García, M.; Codina, L. Los cibermedios y la importancia estratégica del posicionamiento en buscadores (SEO). *Opción* **2016**, *9*, 929–944.
57. Díaz-Noci, J. Cómo los medios afrontan la crisis: Retos, fracasos y oportunidades de la fractura digital. *Prof. Inf.* **2019**, *28*. [CrossRef]
58. Dimoulas, C.; Veglis, A.A.; Kalliris, G. Semantically Enhanced Authoring of Shared Media. In *Encyclopedia of Information Science and Technology*; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA, 2018; pp. 6476–6487.
59. Aitamurto, T.; Aymerich-Franch, L.; Saldivar, J.; Kircos, C.; Sadeghi, Y.; Sakshuwong, S. Examining augmented reality in journalism: Presence, knowledge gain, and perceived visual authenticity. *New Media Soc.* **2020**. [CrossRef]
60. Olsen, R.K.; Solvoll, M.K. Bouncing off the Paywall—Understanding Misalignments between Local Newspaper Value Propositions and Audience Responses. *Int. J. Media Manag.* **2018**, *20*, 174–192. [CrossRef]

future internet

MDPI

Article

An AI-Enabled Framework for Real-Time Generation of News Articles Based on Big EO Data for Disaster Reporting

Maria Tsourma *, Alexandros Zamichos, Efthymios Efthymiadis *, Anastasios Drosou and Dimitrios Tzovaras

Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thessaloniki, Greece; zamihos@iti.gr (A.Z.); drosou@iti.gr (A.D.); dimitrios.tzovaras@iti.gr (D.T.)
* Correspondence: mtsourma@iti.gr (M.T.); efthimios@iti.gr (E.E.)

Abstract: In the field of journalism, the collection and processing of information from different heterogeneous sources are difficult and time-consuming processes. In the context of the theory of journalism 3.0, where multimedia data can be extracted from different sources on the web, the possibility of creating a tool for the exploitation of Earth observation (EO) data, especially images by professionals belonging to the field of journalism, is explored. With the production of massive volumes of EO image data, the problem of their exploitation and dissemination to the public, specifically, by professionals in the media industry, arises. In particular, the exploitation of satellite image data from existing tools is difficult for professionals who are not familiar with image processing. In this scope, this article presents a new innovative platform that automates some of the journalistic practices. This platform includes several mechanisms allowing users to early detect and receive information about breaking news in real-time, retrieve EO Sentinel-2 images upon request for a certain event, and automatically generate a personalized article according to the writing style of the author. Through this platform, the journalists or editors can also make any modifications to the generated article before publishing. This platform is an added-value tool not only for journalists and the media industry but also for freelancers and article writers who use information extracted from EO data in their articles.

Keywords: Web 3.0; article composition; Earth observation (EO); journalism 3.0; media industry; journalistic workflow; journalistic practices; text generation with artificial intelligence (AI); disaster reporting; EarthPress

Citation: Tsourma, M.; Zamichos, A.; Efthymiadis, E.; Drosou, A.; Tzovaras, D. An AI-Enabled Framework for Real-Time Generation of News Articles Based on Big EO Data for Disaster Reporting. *Future Internet* **2021**, *13*, 161. https://doi.org/10.3390/fi13060161

Academic Editor: Gyu Myoung Lee

Received: 12 May 2021
Accepted: 17 June 2021
Published: 19 June 2021

1. Introduction

With today's use of the web, it is possible to access a huge amount of multi-modal information, such as documents, images, videos, etc. This information is generated from different sources and is often difficult to collect. In this new forming environment, while making it possible for journalists to access this huge volume of information, it is often difficult to collect, correlate, and exploit this heterogeneous data, and therefore, it is difficult for journalists to have a complete overview of an issue of a complex topic. This happens because the majority of the information available on the internet is published in an unstructured way. The collection and correlation of unstructured information is a demanding and time-consuming process. Therefore, there is a need for the creation of tools that will collect data from different sources and convert them into structured data with the use of new technologies, allowing all users to have access to them. These new technologies are shaping a new advanced web form called Web 3.0 or the semantic web.

Web 3.0 can be viewed as an extension of the current form of the web. An interpretation of Web 3.0 is that it is an internet service that contains advanced technological features that leverage machine-to-machine interaction but also human-to-human cooperation [1,2]. Web 3.0 entails multiple key elements, such as the semantic web and artificial intelligence (AI) that can be leveraged in order to create added-value web-based applications. Within this context, web-based applications that combine AI and Big Data have gained great value

in recent years and made possible the retrieval of unstructured data from different sources, as well as the analysis and correlation of the extracted information, in order to use them in decision-making or for the automation of time-consuming processes. In addition, the combination of these technologies has led to the analysis of demanding information with great value, such as Earth observation (EO) data.

EO offers a wide and timely covering solution via the fleet of space-borne sensors orbiting the earth at any time and the Big Data piling up at main storage and satellite facilities and hubs of ESA and NASA. The EO data produced are becoming tremendous in volume and quality, and thus, their retrieval, processing, and information extraction is time-consuming and demanding in terms of skills and resources. Therefore, AI methods and cloud technologies have been leveraged for the processing and data extraction of EO data. Once these data are processed, they can offer information of great value that can be leveraged from any end-user without restrictions or need for special knowledge.

One of the areas of interest, where the extracted information of EO data along with the web-based applications of Web 3.0 could be combined, is the media industry and, more specifically, journalism. Corresponding tools of advanced technologies developed in the context of Web 3.0 can change the way journalists collect, process, and eventually interpret information, helping the transition to Journalism 3.0. Journalism 3.0 can be seen as the direction of using advanced technologies by journalists to automate journalistic practices and workflows. Web-based applications using advanced technologies in this context, can be exploited by journalists and automate or semi-automate some of the journalistic practices and workflows followed. In the context of Journalism 3.0, new added-value technologies are applied in the media industry, such as AI, allowing journalists to use the results of scattered information analysis and fusion in order to create news articles promptly for several topics. In particular, new deep learning and machine learning methods offer new capabilities for a variety of automatic tasks, such as automated content production and others, such as data mining, news dissemination, and content optimization [3]. In general, the use of EO data and images, which contain valuable information, could benefit journalists through the provision of additional information extracted from the EO data processing. Such information might concern the percentage of an affected area from a disaster (e.g., flood, fire, earthquake, etc.).

This paper presents the concept of a new innovative web-based platform, called EarthPress. The EarthPress platform aims to deliver added-value products to editors and journalists, allowing them to enrich the content of their publications. EarthPress leverages EO data to generate automatically personalized news articles using also the information extracted from the analysis of EO data. This platform uses AI methods for the aggregation of data from different sources, such as texts from news articles of different websites and posts on social media, and using their analysis and utilization as input for the automatic generation and composition of a news article. The utilization of EO data within the EarthPress platform gives great value to the information that will be provided in the generated text since journalists and users that are interested in the EO data currently cannot easily have access to them or analyze them, because EO data analysis is a demanding and time-consuming procedure.

From the analysis of the collected data, the most trending topics are extracted and provided to the end-users, allowing them to be selected. For each topic, any available information (e.g., EO images, images from social media, social media posts, news articles, etc.) referring to this event are also retrieved and made available to the end-users. With this information, the user can select the ones that s/he wants to use and generate a personalized article through the EarthPress platform. For the personalization of the generated article, the writing style of the author is extracted using AI methods and is transferred to the generated text using previously written articles of the user as a basis. The personalized article that is provided as output of the EarthPress platform is trimmed and finalized by the journalist within the given layout constraints.

The platform includes AI methods for not only the collection and analysis of the data retrieved and the generation of the final article but also for the evaluation of the quality of the retrieved data. Currently, journalists have access to a pile of information for the generation of an article; however, they cannot easily discriminate which of this information is valid and which sources are trustworthy enough. Therefore, in order to avoid spreading misinformation, one of the key aspects of this platform is to provide qualitative data to journalists from verified sources and also to use this verified information for the synthesis of the final news article. Misleading content and fake news included in news articles and posts from social media are eliminated. High quantity journalistic articles and posts from well-known journalist's profiles on social media are collected and promoted. Non-trustworthy sources are excluded from the list of data sources in order to minimize the risk of providing misleading information. The quality of information is the most important feature of the platform as we aim to implement a solution that will have ethical features and will not promote the production of false news or its reproduction. The news article generated by the EarthPress platform will be also evaluated within this scope.

The focus of EarthPress is disaster reporting (e.g., floods, fires, drought, pest, earthquake, erosion/ sedimentation, avalanche), as journalists are challenged in such cases to find the area of interest, collect and analyze any available information in minimal time, and write a detailed article reporting the disaster's status or outcome. The main scope of this platform is to semi-automate the journalistic workflows and provide journalists with the resources and opportunity to create news articles by having a vast amount of qualitative information available. It will act as an intelligent assistant for the journalist, offering enhanced possibilities for content-rich reports. AI techniques make unsupervised querying through Big Data piles feasible. Moreover, EO data resources and geospatial data will become accessible for exploitation by domain-unexperienced users. Tailor-made solutions will be promoted. Hence, the editor's or journalist's profile will be consulted to allow for a high-level specification of learning AI tasks and reporting.

1.1. Related Work

An important parameter related to the functionality of the platform includes the analysis of EO data and, specifically, EO images. A variety of such platforms exist [4,5], most of which offer tools for the processing and analysis of data from satellites. The platform will offer the feature for automatic processing of EO images in real time.

Global Disaster Alert and Coordination System (GDACS) [6] provides a disaster manager and disaster information systems worldwide and aims to fill information and coordination gaps in the aftermath of major disasters in the first phase. The platform is a cooperative framework of the United Nations and the European Commission. The GDACS provides real-time access to web-based disaster information systems and related coordination tools.

Emergency and Disaster Information Service (EDIS) [7] is another platform that aims to monitor and document all the events on Earth that may cause disaster or emergency. The platform is operated by the National Association of Radio Distress-Signaling and Infocommunications, and it monitors and processes several foreign organizations' data to get quick and certified information.

GLIDE [8] is another platform that provides extensive access to disaster information in one place, as well as a globally common unique ID code for disasters. Documents and data pertaining to specific events can be retrieved from various sources or are linked together using the unique GLIDE numbers. In addition to a list of disaster events, GLIDE can also generate summaries of disasters, e.g., by type, country, or year and also provides export of data sets in charts, tabular reports, or statistics.

Different methods for the processing of EO images were used, according to literature, and particular methods were applied for building footprint extraction from EO images. For this reason, architectures for image segmentation [9–11] have been reviewed. Other

methods of processing of EO images used for the implementation of the platform concern the detection of the affected areas from floods and water in general [12,13] and fire [14,15].

One of the important features supported by the platform concerns the personalization of the generated text. For transferring users' writing style according to literature, methods to extract the user's sentiment [16,17] have been implemented.

In literature, the task of text generation includes a vast number of methods. A summary of multiple texts that maintains the information of these texts can be used for text synthesis. State-of-the-art approaches in the field of summarization, in general, and multi-document summarization, in particular, are all based on the Transformer [18], a model that leverages the mechanism of attention and achieves avoidance of utilizing recurrency and convolution [19–22]. The state-of-the-art PEGASUS model [23] is a large Transformer-based encoder–decoder model pre-trained on massive text corpora with a self-supervised objective and is designed for use in both single and multi-document summarization cases. Notably, some other, more general purpose state-of-the-art NLP models, such as BART [24] and T5 [25], can produce results comparable with those of PEGASUS, in a few-shot and zero-shot multi-document summarization settings, suggesting that unlike single-document summarization, highly abstractive multi-document summarization remains a challenge [23].

Data-to-text generation (D2T) is a subtask for the generation of text using piles of data as input. Data-to-text generation is a promising subtask of study with many applications in media, paper generation, storytelling, etc. In their work, Harkous, H et al. [26] present the DATATUNER, a neural end-to-end data-to-text generation system that makes minimal assumptions about the data representation and target domain. DATATUNER achieves state-of-the-art results in various datasets. Another work presented by Kanerva, J et al. [27] aims to generate news articles about Finnish sports news using structured templates/tables of data and pointer-generation network. In the work of Rebuffel, C et al. [28], a hierarchical encoder–decoder model is proposed for transcribing structured data into natural language descriptions. In other papers, like the one presented by Mihir Kale [29], they use pre-trained models, such as T5 [25] and BART [24], for data-to-text generation, achieving state-of-the-art results in various datasets.

1.2. Motivation and Contribution

In the era of Big Data, where a vast amount of data is produced on a daily basis, there is the necessity for the implementation of new tools or platforms that are able to facilitate the selective collection, filtering, and processing of this information so useful information can be extracted and used in different fields of the market. One of these fields is the media industry. Journalism is closely related to the retrieval, evaluation, and filtering of raw information, as well as, to the combination of knowledge retrieved from multiple sources and the publishing of news stories that are based on these data. Thus, Web 3.0 applications that are able to facilitate the above journalistic practices are of great significance and comply with the transition towards Journalism 3.0.

In this work, we present EarthPress, an interactive web-platform whose scope is to facilitate journalists and editors in synthesizing news articles about disasters. The platform combines various methods from different scientific fields, such as information retrieval and artificial intelligence and, more specifically, natural language processing and image processing. EarthPress services allow its users to have access to multimedia data (e.g., documents, videos, images, etc.) from multiple sources regarding disasters such as floods and fires. Moreover, the EarthPress platform detects automatically, and in real-time, breaking news related to disasters. The latter is achieved by monitoring publications in social media. The detected breaking news is presented to journalists through the platform's interactive user interface (UI). Additionally, the platform consists of models that are able to retrieve, process, and extract useful information and statistics from EO data, which can be used by editors and journalists in order to enrich the content of their publications. Finally,

the platform provides services that can combine all the above and generate automatic ready-to-print news articles, considering also the writing style of the journalist.

In summary, the platform can facilitate the following journalistic practices:

- Data collection: raw data need to be available (search for data on the web).
- Data filtering: the process of filtering relevant information from the news story.
- Data visualization: the process of transforming data and creating visualizations to help readers understand the meaning behind the data.
- Story generation (publishing): process of creating the story and attaching data and visuals to the story.

2. Materials and Methods

This section presents the architecture of the platform, along with a description of the modules and sub-modules included in it. Initially, and before the presentation of the plat-form's architecture, the method for collecting user requirements and specifications in order to define the functionalities of the platform is presented.

2.1. Collection of User Requirements

In software development, an important step before the implementation of a platform is the collection and analysis of user requirements. Thus, before the implementation of the EarthPress platform and the definition of its architecture, an online international workshop was conducted in March of 2021, aiming to determine the needs of journalists, current practices followed, and attitudes towards the EarthPress platform. The target audience, 134 participants, of the workshop was both journalists and academics. Many of the participants had experience in using EO data and in using other platforms for article generation. In the workshop, the use and importance of EO data were presented, along with the concept of the EarthPress platform. For this workshop, a questionnaire including both closed and open-ended questions aiming to identify the opinion, needs, and current practices followed by the participants when writing an article was designed. The questionnaire was available online for one week, and the link was sent to all participants at the end of the workshop. The results of the collected data analysis are presented in this sub-section.

The designed questionnaire included two parts. In the first part, demographic information of the users, such as their gender, age, background, etc., was requested (Table 1). Additionally, one of the questions in the first part asked the users to answer if they previously used or are currently using information from EO data in their articles and if they have previously used other platforms for article generation. This was general information requested in order to have a better view of the sample used.

Table 1. Demographic information of the Participants.

Question	Answer-Distribution
Gender	Male (19.4%), Female (79.1%), Prefer not to say (1.5%)
Age	19–21(82.6%), 22–51(17.4%)
Education	Students (95.5%), Other (4.5%)
Experience in using EO data	Yes (26.1%), No (73.9%)
Experience in using article generation platforms	Yes (7.5%), No (92.5%)

The second part of the questionnaire concerned the retrieval of information about the type of data and current practices used during the generation of an article and the flow that is followed for the retrieval and analysis of EO data. In this part, the participants highlighted the main problems or issues currently existing in the field of journalism related with the scope of the platform. The issues presented by the majority of the participants concerned the fact that they did not have the means to collect adequate information for a certain topic and that the collection of all information for the generation of an article was a

time-consuming task. As a second issue, they mentioned the restricted time frame they had for writing an article. Additionally, most participants responded that it was difficult to collect EO images related to a specific area of interest where the event had occurred and analyze them in order to extract any data included. Regarding the usefulness of the EO data, they mentioned that they would be interested in using such information when creating an article; however, this information should be processed. Moreover, another answer provided by some of the participants highlighted the existence of fake news among the news stories published by other journalists for an event as a problem.

Additionally, the participants mentioned that it was crucial for the journalists to receive this information in a timely fashion and have access to all available multimedia and data (e.g., videos, analytics, local news posted, etc.) on a specific topic in order to use them for the article composition. In this context, the EarthPress platform will provide images and local news retrieved for a certain event/disaster to the end-users, allowing them to select the information that is taken into consideration for the text generation.

From the open-ended questions, the suggestions proposed mentioned that the platform would be a useful tool for journalists and other professionals and that by using this platform, journalists would have a chance to be more precise, avoid fake news, and provide qualitative news articles to the public. The elimination of misleading information and data sources was mentioned in more than one answer provided, making the importance of this issue a high priority for the EarthPress platform.

The requirements and specifications extracted from this survey were valuable for the definition of the EarthPress platform's architecture and functionality. Initially, the fact that users presented the text generation process as a time-consuming task and that EO data were not easily retrievable on time was very important, since these two major issues are the basis of the platform's scope. The platform strives to solve the aforementioned issues through the early detection of trending issues, the presentation of qualitative data to the user, and the direct extraction of information produced by the analysis of EO data. Furthermore, it is important that all information that is currently useful for the journalists during the composition of an article, such as posts from social media platforms and articles from websites, is available to the users through a platform in order to be easily accessible at all times.

Considering the generation of an article and based on the fact that most of the participants had previous experience with such platforms, the requirement extracted concerns about the ability of a user to edit an article and receive qualitative information extracted by the analysis of EO data on time in order to use them for the information extraction. Regarding the multimedia used for the final article, images, and statistics, along with video, were highly rated options by the participants. Therefore, the platform should be able not only to retrieve this information when it is available but also to provide it to the users and allow them to select which information will be used for the final article.

According to the analysis of the data provided in this sub-section, the architecture of the EarthPress platform was enhanced in order to include not only the basic modules but also to include the corresponding modules and meet the requirements extracted.

2.2. System Architecture

The system's architecture is presented in Figure 1. The proposed architecture consists of eight main modules, which are the: (a) user interface, (b) user manager, (c) breaking news detector, (d) data fusion (e) database manager (f) image processing, (g) journalist profile extractor, and (h) EarthBot module. The architecture of the platform includes two parts, the front-end, and the back-end. The front-end concerns the design of the platform's user interface (UI), while the back-end includes the platform's database, the implemented AI methods for image processing and text synthesis, as well as, algorithms for data collection and data filtering. A detailed description for each sub-module is provided in the following sub-sections, while a description of the user interface is available in Section 3.

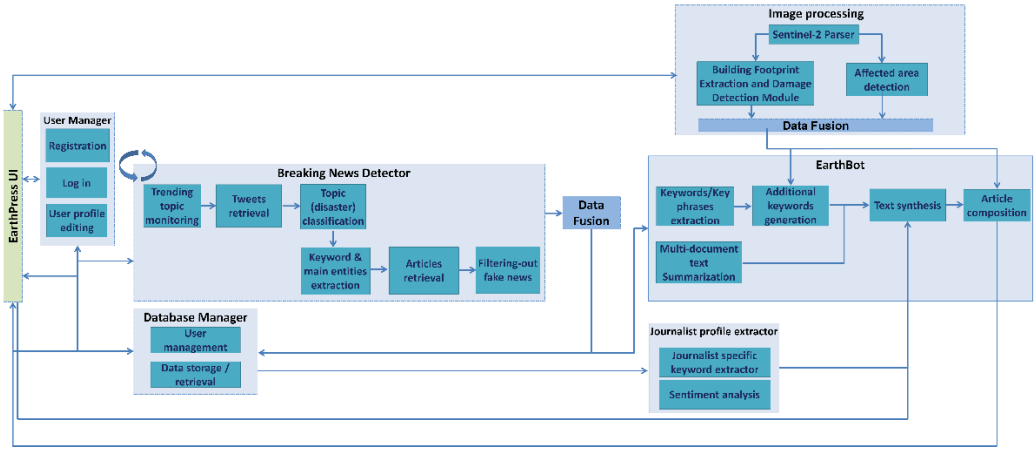


Figure 1. Platform's architecture.

2.3. User Interface

The user interface module refers to the interactive interface of the platform, which allows users to interact with the other modules of the system, as well as visualize the results of the user's searches and the generated article. The purpose of this module is to allow the smooth interaction of the users with the system. Through this interface, users will be able to interact with the system in a user-friendly way without any restrictions or requirements.

2.4. User Manager

The main scope of this module is the registration of new users to the EarthPress platform, as well as the logging-in of the already registered users. It is responsible for (a) the registration of new users to the platform, (b) the authentication process of users when they log in to the platform, and (c) the updating of the users' profile. The corresponding APIs are included in this module, aiming at the exchange of all data required among the related modules.

2.5. Database Manager

This module is a data storage mechanism for storing all the available information from all modules of the platform. The types of data stored vary according to the needs of each module. The input data for this module are outputs from other modules, such as tweets, articles, and any additional information provided by the user via the user interface, such as any previously written documents by the user that will be used for the personalization of the generated text.

2.6. Breaking News Detector

The instant detection of breaking news, as well as the collection of data related to the breaking news, is of great significance for journalists. In this scope, a module that detects breaking news and collects relevant news data is included within the platform's architecture. Within the EarthPress platform, the important news that we are trying to detect is related to either natural or human-made disasters. To detect breaking news related to disasters, the platform uses the Twitter social network.

The breaking news detector is composed of several sub-modules that perform different functionalities. Two of the sub-modules are related to the data retrieval from Twitter. At first, the trending topic-monitoring sub-module downloads tweets that contain certain, pre-defined, keywords related to disasters, and second, the topic (disaster) classification sub-module monitors the trending topics of certain areas and recognizes trending topics related to disasters.

From the tweets that have been extracted and are related to a natural disaster, it is important to extract information such as the type of disaster, the location, and the date of the event. For this reason, the keywords and main entities extraction sub-module is used to extract the most important words and phrases from the posts retrieved, along with the aforementioned information. For this task, the YAKE algorithm has been used [30]. These keywords are used as input by the article retrieval sub-module, whose aim is to collect news articles that will contain these certain keywords.

Finally, to ensure that the output of the breaking news detector does not contain articles and tweets with false/misleading content, the collected information is checked for their reliability and the detection of fake news. It is essential to ensure that the output information of the breaking news detector does not contain articles and tweets with false/misleading content, as these data will be used to compose the final news article in the EarthBot module. For this reason, deep learning methods have been developed that automatically detect and filter out any text with misleading content to ensure the quality of the output information.

2.7. Data Fusion

As mentioned previously, the information that will be used for the generation of the final article comes from different sources. This information needs to be correlated and stored in a grouped way in the database so that it can be used more easily when requested. To store the information in a grouped way so that it refers to common events for every user in the database, a mechanism for comparing and correlating the different information has been implemented.

2.8. EO Image Processing

The scope of this module is to analyze EO images and extract any available information included. The EO images are retrieved using a sentinel-2 parser responsible for downloading sentinel-2 images from the Copernicus Open Access Hub, according to an area of interest and a timestamp. The users select these input parameters when they select a topic of interest from the user interface presented in Section 3. This image is further processed for the detection of affected areas and changes in buildings that have occurred from a disaster. Two distinct sub-modules with different scopes have been included in the EO image processing module: the first one, called the "building footprint extraction and damage detection sub-module" aims to detect damages within an urban area using RGB images and is based on the buildings that exist within a certain area. The second one. Called the "affected area detection" sub-module, is independent of the type of area and can detect floods and burned areas based on sentinel-2 images.

The outcomes of the processing are the processed EO images retrieved, accompanied by rough statistics on the impacted area (e.g., surface, type of land cover/land use affected). The numerical information, provided as output, is taken into consideration during the generation of the news article, giving an added value to the generated news article. On the other hand, the processed EO images are made available to the end-user through the User Interface, allowing them to select if they would like these images to be presented in the final article or not.

The sub-modules included in this module are:

2.8.1. Sentinel-2 Parser

The sentinel-2 parser has been implemented for the retrieval of EO images. This parser receives input coordinates showing the location of the event selected by the user and a timestamp. A before and an after depiction of the location provided are retrieved from the Copernicus Open Access Hub in order to be processed through the sub-modules included in the image processing module.

This sub-module is not only responsible for the retrieval of the sentinel-2 images but also for their conversion to RGB images in order to be used as input in the building

footprint extraction and damage detection sub-module. It should be mentioned that the sentinel-2 data that will be retrieved will include 13 spectral bands: four bands at 10 m, six bands at 20 m, and three bands at 60 m spatial resolution.

2.8.2. Building Footprint Extraction and Damage Detection Sub-Module

This sub-module implements a key EO data processing function by providing the capability to detect buildings in satellite imagery and the capability to detect changes in the footprint of the building(s) given images from a time point preceding and a time point following a (disastrous) event (e.g., an earthquake).

The building footprint extraction and damage detection sub-module uses RGB image(s) as input and provides as output a PNG image file of the segmentation map depicting the two classes, "building" and "not building", along with a text file presenting the building footprint area in square meters. Both sets of data will be used for the synthesis of the final article.

2.8.3. Affected Area Detection

The sub-module of affected area detection aims to detect the affected areas using the images retrieved from the sentinel-2 parser as input. Using two different methods, this sub-module can detect both the areas affected by water, through the water change detector and the burned areas, through the burned area detector. The water change detector processes images from the sentinel-2 and outputs processed images and numerical data relating to flood disasters. Similarly, the burned area detector also processes images from the sentinel-2 sub-module and extracts processed images and numerical data relating to fire disasters.

Both modules provide the processed images as output, depicting the affected areas with the use of additional layers over the initial image and numerical data including the percentage of the affected area. The images and the numerical data resulting from this analysis are used for the composition of the final article. The images can be included in the final article, while the numerical data will be used within the context of the generated text, aiming to provide additional information to the user, which in other cases might not be feasible to be retrieved.

2.9. Journalist Profile Extractor

The scope of the EarthPress platform is not only to generate articles that include EO data but also to provide personalized articles that follow the writing style of each author. Each author has a unique way of writing and presenting news. In order to personalize the generated text and, in particular, to transfer the style of the writer to the generated text, a module that creates the user's profile from their previously published articles has been designed. This module is intended to define the user's profile by extracting the words/phrases that a journalist uses more frequently on his/her articles, along with the sentiment that the journalist usually uses when composing an article on a certain subject. This module consists of the following submodules:

2.9.1. Journalist Specific Keyword Extractor

As mentioned previously, the aim of the journalist specific keyword extractor is to extract the most significant and more frequently used keywords from the documents provided by the user when creating their profile. The most frequent words that a user tends to use in the document are extracted using the keywords extraction method, similar to those used in Section 2.6. These words, named keywords in this context, are an important asset for the article generation as with the words used by the user in an article, the dynamics of the text are alternated.

2.9.2. Sentiment Analysis

The sentiment is an important asset for writing an article. Journalists, and authors in general, tend to write an article having a certain sentiment according to the event's outcome. The tone of each article can be retrieved by analysis of a document, resulting in the extraction of the document's sentiment. This characteristic is important for the personalization of the article through text style transfer since the tone of the article will be adjusted to the sentiment generally used by the author in similar articles.

Using the database manager, the documents provided by the user as input during the user's profile creation are retrieved from the database. These documents are further analyzed in the sentiment analysis in order to extract the sentiment used by the user in the articles. For this purpose, the BERT [16] model has been developed and integrated within the platform. For each journalist, his/her previously published articles are considered. The sentiment that is more frequently used by a journalist for describing a news story related to a certain type of disaster will be finally extracted and passed to the EarthBot module for the final article creation process.

2.10. EarthBot

The EarthBot module is the last module of the EarthPress architecture and is responsible for the automatic generation of news articles. The EarthBot is responsible both for the text synthesis of the article as well as the final article composition that integrates both the generated text and multimedia content selected by the user. In order to achieve that, EarthBot integrates different AI technologies, such as multi-document summarization, name entity recognition and keywords extraction, and data-to-text generation, for the generation of the textual part of the article.

The input data of EarthBot are posts from social media and news articles relevant to a disaster, along with statistics that are extracted from the processing of EO data by the EO image-processing module. The social media posts and the news articles are processed, so the most important keywords and main entities can be extracted from them. Additionally, these posts are inserted into the multi-document summarization method supported in order to have a common summary of all documents as output. These two types of data are passed as input to the news article text synthesis sub-module, which is responsible for generating the text content of the final article. This sub-module supports two options that are further described in sub-Section 2.10.3.

Moreover, the news article generation sub-module receives the journalist's profile as its input for the personalization of the generated text. The composition of the final article is achieved by combining the produced text and images selected by the user by the article composition sub-module. In order to achieve all the aforementioned tasks, this module consists of the following sub-modules:

2.10.1. Additional Keywords Generation

The terms keyword and key phrases are used interchangeably in this context. Apart from the extraction of the keywords that are present in a text (i.e., in posts and documents retrieved), additional, absent keywords that are not present in the initial text can potentially be generated using deep learning sequence-to-sequence models. These additional keywords are used as input for the text synthesis sub-module. For the implementation of the additional keywords generation, the method described in [22] is used.

2.10.2. Multi-Document Text Summarization

Multi-document summarization is the task of producing summaries from collections of thematically related documents, as opposed to the single-document summarization that generates the summary from a single document. In the scope of the EarthPress platform, for the text generation task, a multi-document summarization method is used initially to generate summaries of groups of thematically related articles or tweets. For the implementation of multi-document summarization, the PEGASUS [24] model has been

developed and integrated within the platform. The produced summary will be further provided as input for the text generation.

2.10.3. News Article Text Synthesis

This sub-module implements the task of generating the textual part of the news article. The generated article describes the event that the user chooses from the available topics resulting from the breaking news detector. The generated text should be consistent, coherent, and syntactically correct. Once the text is generated, the user can edit it and make changes or additions according to his/her preferences.

The user has the opportunity to choose one of the two options supported for the generation of the news article. For the first option, all available information and data extracted from the EO data processing are mapped to predefined text templates. For this reason, a rule-based system will be used to translate the data to map them to the templates. For the implementation of these methods, the method described in [31] is used.

For the second option, two deep learning approaches are followed. For the first approach, a deep neural network is deployed. This method uses data keywords/key-phrases and data extracted from the EO image processing module and the journalist profile extractor as input. The processing of the data is based on a tokenizer that encodes the data into useful representations for computational purposes. Finally, the processed data are used as input to the model used for the text generation. For this approach, the T5 model [25] is used.

For the second approach, a deep neural network is implemented that uses a summary from different textual data as it was produced by the multi-document summarization sub-module. Subsequently, the output summary is aggregated with information from the EO image-processing module and the journalist profile extractor and forms the final news article.

A key difference between the two approaches lies in the different nature of the input data for the deep learning models. In the first approach, keywords/key-phrases and other non-textual data extracted from tweets and articles will be used as input data. In the second method, the tweets and articles themselves will be used for input data without an intermediate stage of extracting information from them. The platform will support both approaches.

2.10.4. Article Composition

This submodule is responsible for the composition of all the individual elements that make up the final article. Data from different modalities (text, images) should correspond to each other (e.g., the content of the first paragraph of the text should correspond with the following image, captions should correspond with the images). If the user chooses the first method to produce the text, as described in Section 2.10.3, then the texts and the corresponding images will be placed in specific predefined positions.

If the user chooses the second method for the final news article generation, a heuristic algorithm is used to correlate data from different modalities and synthesize a well-presented article. The end-users will be able to edit the final article (e.g., rearrange the text and the image places) according to their own needs and preferences. The users will be able to make changes in the content of the generated text. The input data for this submodule are the generated news article, processed EO images resulted from the image processing module, links, hashtags, and image captions.

3. System Presentation

The EarthPress platform is a web-based application for Web 3.0, allowing users to receive information about events occurring in a certain location, retrieve analyzed EO data, and generate a news article by selecting one of the two options supported for the text generation. This platform provides a user-friendly interface through which the users can

interact with the EarthPress platform and its functionalities. In particular, a user through the user interface of the EarthPress platform has the following abilities:

- Log in and register to the platform;
- Create and edit their profile and upload any previous articles that they want to be used for text generation;
- Receive a list of current breaking news, filter and select a category;
- View the retrieved EO images and the features extracted from them;
- Generate an article, edit, and store it.

The user can fill in the profile when s/he has already registered in the system. The user can set the language that s/he uses when writing an article. Moreover, they can upload their previously published news articles (Figure 2) in order for the system to recognize and extract their writing style. The latter will be used as input to the module that is responsible for the text synthesis (EarthBot) in order to personalize the automatically generated article. The module that is responsible for the journalist profile extraction is described in Section 2.8. The EarthBot module is described in Section 2.9. The final article will be composed of the automatically personalized generated text, along with any retrieved images selected from the user. The images that will be included in the final article will be both processed EO images and images retrieved from Twitter's posts.

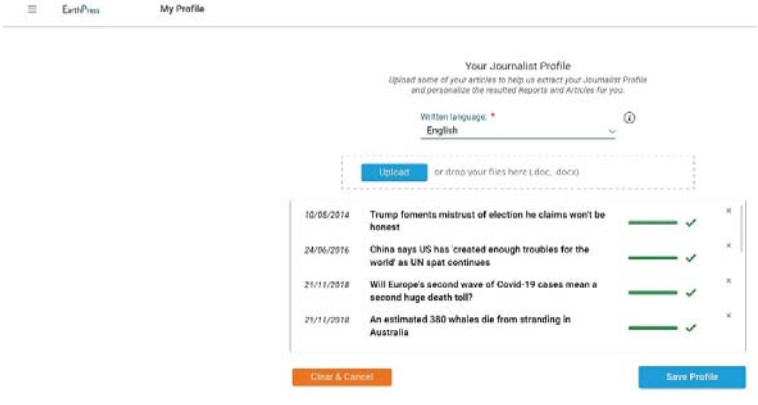


Figure 2. User profile view, depicting the ability of a user to upload and remove any uploaded documents stored in his/her profile.

When the user has created his/her profile, s/he is transferred to the main view of the platform. The main view includes initially a table presenting a list of the breaking news events, along with additional information (Figure 3) resulting from the breaking news detector module, presented in Section 2.5. This table is updated periodically aiming to provide any new information to the user. The user is allowed to interact with this table and filter this list of breaking news according to the features provided for each event, such as the location of an event, etc.

The user is able to be select the area of interest from the interactive map. The next step is to click on the corresponding button in the main view and receive related information about the selected event. Based on the requirements presented in Section 2.1, users tend to search on social media platforms for information about a topic; they are using such information for the creation of an article, along with other news articles and images posted on social media. Therefore, all this information should be available to the user on the related material section in the user interface (Figure 3). The user should be able to look at all the available material and select the articles and posts that s/he wants to be included in the final article.

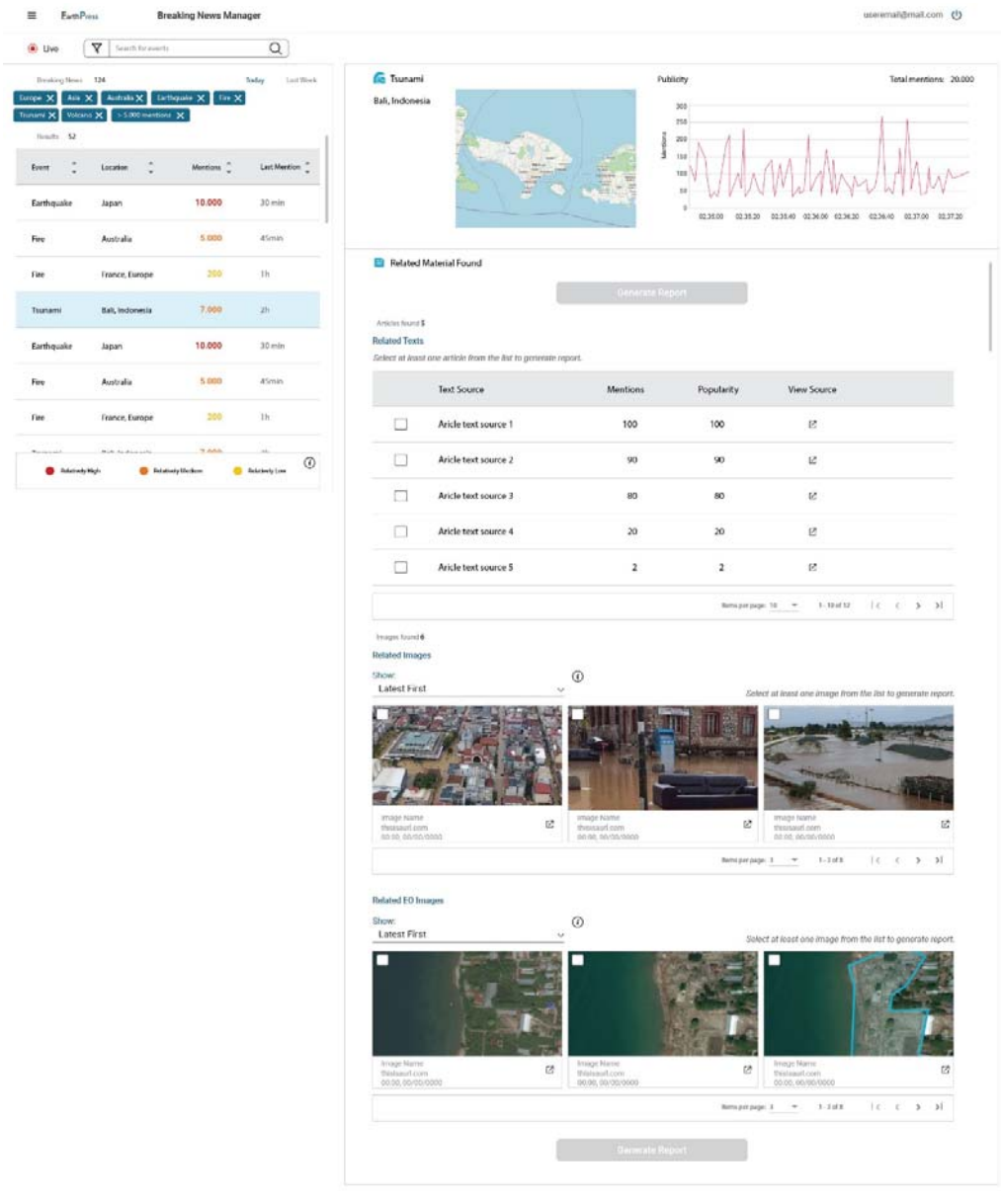


Figure 3. Visualization of all the available material and EO images based on the selected event.

As presented in Figure 3, the user can select the posts and articles that will be used for the generation of the final article. Additionally, the user can select from two different categories of multimedia. The first category, called "related images" in Figure 3, includes images retrieved from social media and are the ones posted by users. The second cate-gory, called "related EO images" in Figure 3, includes the EO images retrieved and the processed ones resulting from the image processing module. From the related articles, posts, and images provided, users can select at least one of them to be used in the final article.

Once the user has selected the images that s/he wants to be included in the final article and at least one of the related articles and posts provided, s/he can proceed to the generation of the article. Aiming to increase user experience, a new view is created (Figure 4), allowing the user to view, edit, store, and update the generated article. In particular, through this view the user can remove the keywords that were initially used for the generation of the text, include new keywords from the list provided and re-generate a text and view the generated article. An important element for the generation of the text is the order in which the keywords are entered so that there can be coherence and flow in the generated text. Therefore, the user will be able to re-arrange the order of keywords. In this list, additional keywords can be added. The additional keywords can be retrieved from another list of keywords extracted from the corresponding modules presented in Section 2.9.1.

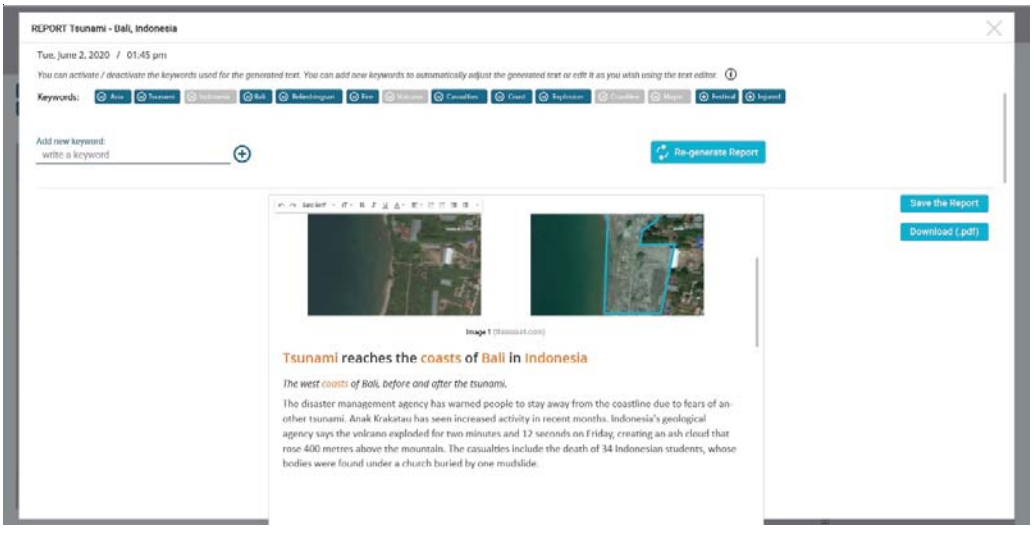


Figure 4. Additional view presenting the automatically personalized generated text.

This final article will be trimmed and finalized by the users themselves within the given layout constraints. Furthermore, the user is allowed through this view to save the generated article on his/her user profile. In addition, the final article will be downloaded upon request through this view.

4. Discussion

In recent years, the amount of data (documents, images, videos, etc.) that are produced daily has increased tremendously. This has raised many challenges in different fields, such as journalism, where the development of new methods that are able to collect and extract useful information from these vast amounts of information is necessary. Journalistic practices, such as data-driven journalism, follow four basic steps toward the creation of a news story, which are: (a) the data collection, (b) data filtering, (c) data visualization, and (d) story generation. Data collection refers to the collection of raw data while data filtering refers to the process of filtering and maintaining only the information relevant to the news story. Data visualization is the process of transforming data in such a way and creating visualizations to help readers to understand the meaning behind the data. Finally, story generation (publishing) refers to the process of creating the story and attaching data and visuals to the story.

However, in each of the aforementioned steps, several difficulties are raised. Multimedia data collection can be laborious due to the huge amount of available data and to the

multiple data sources. Moreover, in many cases, programming knowledge may be required in order to collect data from certain sources. Additionally, the collected data may come in different specialized formats and may be difficult to be checked and filtered from a person or a group of persons. Furthermore, much of the collected information may refer to fake news that has to be filtered out. In addition, the extraction of additional knowledge from data, such as the EO data, and the creation of easily understandable visualizations may require expert knowledge. Finally, the combination of the available data and the limited response time for publishing breaking news makes it a very challenging task. The whole process is a time-consuming procedure that requires the development of systems that can tackle the above difficulties.

Towards this scope, this paper presents an innovative Web 3.0 platform, called EarthPress, aiming to facilitate professionals of the Media industry by automating many of their journalistic practices. More specifically, the platform is intended to act as a supportive tool in many steps of the journalistic workflow, from data collection and data filtering to the extraction of useful information from the collected data and the automatic synthesis of news stories. In addition to the provided services, the platform aims to deliver value-added products to the editors and journalists based on EO data, allowing them, thus, to enrich their publications and news articles. Such information is important in the news that is related to disasters such as floods and fires.

EO data are freely available in large quantities. However, the main obstacle to their wide use by journalists relates to the difficulties in accessing and even processing them so as to extract additional meaningful information without expert knowledge. Through this platform, users have the possibility to retrieve and extract valuable information from EO data and, specifically, from EO images, without having prior knowledge of processing satellite images. The collected data are automatically processed and provided to the journalists as valuable information regarding the effect of a disaster on a certain area.

The major market segments that EarthPress targets are: (a) local newspapers, which are interested in providing breaking news with a high level of personalization; (b) nationwide newspapers, which are more interested in producing worldwide information; and (c) ePress, which are generally more specialized and could be interested in accessing more elaborated information regarding disasters and EO data.

In this paper, the platform's architecture and its components were presented and analyzed, as well as, the requirements received by journalists and the platform's user inter-face. The basic services provided by the platform are: (a) the detection of breaking news related to disasters, (b) the collection of EO data from Copernicus and multi-media data from social media and news sites, (c) the filtering of the collected data based on their relevance with a disaster event and their credibility, (d) the extraction of useful information related to a disaster and its effect on an area from EO data through the use of image processing techniques and (e) the automatic text synthesis of news stories through the use of AI models that are personalized according to the writing style of each journalist. Each of these services includes many challenging tasks, from multisource and multi-media data acquisition and filtering to the fake news detection and the processing of the EO data, as well as the generation of personalized news articles. The latter deals with various challenging tasks of the NLP field, such as the extraction of the most important keywords from a given text, the writing style transference for the personalization of the generated text, and the text generation of the news article. Text generation is an open research topic whose scope is to provide human-like written text that is coherent and meaningful. The training of AI models that deal with text generation is usually a quite demanding process that requires resources, a huge corpus of training data, and many hours of training. In the scope of EarthPress, the generated text should be also relative to the news story's topic and should avoid containing misleading or erroneous information.

All the aforementioned services included within the integrated platform of EarthPress can facilitate the journalists significantly in order to reduce the needed time for collecting, evaluating, filtering, combining, and finally publishing news stories ready-to-be printed.

However, there are some limitations in the proposed platform that should be taken into consideration. As it was mentioned previously, text generation is a challenging NLP task of great research interest. The final generated text should be paid attention to in order to avoid including misleading or false information, while it should be also coherent and relevant to the news topic. The fake news detector should check the generated text so the credibility of the generated text can be ensured. Moreover, the retrieval and processing of EO data may be a time-consuming process that may require some hours of processing. Additionally, the accuracy of the results may vary depending on the characteristics of the available EO data (e.g., different Spatio-temporal resolutions, cloud cover, etc.).

5. Conclusions

With the advent of Big Data, the need for innovative tools that are able to handle this vast amount of data and extract information of added value has emerged. In the field of journalism, the latest journalistic practices require the automatic collection and filtering of information related to a news story from multiple sources, as well as the automation of the procedure of the news stories generation.

In this work, the concept of the EarthPress platform was presented, an assistive Web 3.0 platform that targets to facilitate professionals of the media sector in each step of the journalistic procedure. The EarthPress platform intends to facilitate journalists in collecting data and automatically composing articles related to disasters such as floods and fires. Moreover, it aims at extracting added value information from EO data that can be included in the final news story. Following this, in this paper, the system's architecture of EarthPress is presented, which consists of several modules and sub-modules, each of which serves a different purpose. The main modules of the EarthPress as they are presented in Figure 1 are, namely, (a) the user interface, (b) the user manager, (c) the database manager, (d) the breaking news detector, (e) the data fusion, (f) the EO image processing, (g) the journalist profile extractor, and (h) the EarthBot.

For the definition of the architecture and the user interface of EarthPress, a workshop with more than 130 participants of the media sector was conducted, aiming to identify the requirements and the specification of the EarthPress platform. The collected requirements indicated that a tool or a platform that can facilitate the journalistic procedure in its different steps is of great interest to the media sector. Additionally, the presented platform will include AI methods for the validation of the collected data to avoid using misleading information for the generation of the final news article.

The future steps regarding EarthPress are the implementation and the testing of each of the architecture's modules, as well as the evaluation of the platform in real-time with real users of the media sector.

Author Contributions: Conceptualization, A.D.; methodology, A.D., A.Z., M.T., and E.E.; writing—original draft preparation, E.E., A.Z., and M.T.; writing—review and editing, E.E., A.Z., and M.T.; supervision, A.D.; project administration, D.T.; funding acquisition, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by grants from Horizon 2020, the European Union's Programme for Research and Innovation under grant agreement No. 870373-SnapEarth. This paper reflects only the authors' view, and the Commission is not responsible for any use that may be made of the information it contains.

Data Availability Statement: Not Applicable, the study does not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pileggi, S.F.; Fernandez-Llatas, C.; Traver, V. When the Social Meets the Semantic: Social Semantic Web or Web 2.5. *Future Internet* **2012**, *4*, 852–864. [CrossRef]
2. Fuchs, C.; Hofkirchner, W.; Schafranek, M.; Raffl, C.; Sandoval, M.; Bichler, R.M. Theoretical Foundations of the Web: Cognition, Communication, and Co-Operation. Towards an Understanding of Web 1.0, 2.0, 3.0. *Future Internet* **2010**, *2*, 41–59. [CrossRef]

3. Kotenidis, E.; Veglis, A. Algorithmic Journalism—Current Applications and Future Perspectives. *J. Media* **2021**, *2*, 244–257. [CrossRef]
4. Gomes, V.; Queiroz, G.; Ferreira, K. An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sens.* **2020**, *12*, 1253. [CrossRef]
5. Romeo, A.; Pinto, S.; Loekken, S.; Marin, A. Cloud Based Earth Observation Data Exploitation Platforms. Available online: https://ui.adsabs.harvard.edu/abs/2017AGUFMIN21F..03R/ (accessed on 17 June 2021).
6. De Groeve, T.; Vernaccini, L.; Annunziato, A.; Van de Walle, B.; Turoff, M. Global disaster alert and coordination system. In Proceedings of the 3rd International ISCRAM Conference, Brussels, Belgium, 1 January 2006.
7. Lee, J.; Niko, D.L.; Hwang, H.; Park, M.; Kim, C. A GIS-based Design for a Smartphone Disaster Information Service Application. In Proceedings of the 2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering, Jeju, Korea, 23–25 May 2011; pp. 338–341.
8. Nishikawa, M.S. GLobal Unique Disaster IDEntifier Number (GLIDE): For Effective Disaster Information Sharing and Management. In Proceedings of the International Conference on Total Disaster Risk Management 2003, New York City, NY, USA, 2–4 December 2003; Volume 2, p. 4.
9. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
10. Gupta, R.; Shah, M. Rescuenet: Joint building segmentation and damage assessment from satellite imagery. *arXiv* **2020**, arXiv:2004.07312.
11. Gupta, R.; Goodman, B.; Patel, N.; Hosfelt, R.; Sajeev, S.; Heim, E.; Doshi, J.; Lucas, K.; Choset, H.; Gaston, M. Creating xBD: A dataset for assessing building damage from satellite imagery. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 10–17.
12. Kordelas, G.A.; Manakos, I.; Aragonés, D.; Díaz-Delgado, R.; Bustamante, J. Fast and Automatic Data-Driven Thresholding for Inundation Mapping with Sentinel-2 Data. *Remote Sens.* **2018**, *10*, 910. [CrossRef]
13. Kordelas, G.A.; Manakos, I.; Lefebvre, G.; Poulin, B. Automatic Inundation Mapping Using Sentinel-2 Data Applicable to Both Camargue and Doñana Biosphere Reserves. *Remote Sens.* **2019**, *11*, 2251. [CrossRef]
14. Parsons, A.; Robichaud, P.; Lewis, S.; Napper, C. Field Guide for Mapping Post-Fire Soil Burn Severity. Available online: https://www.fs.fed.us/rm/pubs/rmrs_gtr243.pdf (accessed on 17 June 2021).
15. Emergency Mapping Guidelines. UN-SPIDER. 2018. Available online: https://www.un-spider.org/sites/default/files/IWG_SEM__Guidelines_Fire_chapter_SERTIT_2_0.pdf (accessed on 17 June 2021).
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
17. Gao, Z.; Feng, A.; Song, X.; Wu, X. Target-Dependent Sentiment Classification with BERT. *IEEE Access* **2019**, *7*, 154290–154299. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
19. Liu, Y.; Lapata, M. Hierarchical transformers for multi-document summarization. In Proceedings of the ACL 2019—57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5070–5081. [CrossRef]
20. Goodwin, T.R.; Savery, M.E.; Demner-Fushman, D. Flight of the PEGASUS? Comparing Transformers on Few-Shot and Zero-Shot Multi-document Abstractive Summarization. In Proceedings of the COLING—International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; Volume 2020, p. 5640.
21. Liu, P.J.; Saleh, M.; Pot, E.; Goodrich, B.; Sepassi, R.; Kaiser, Ł.; Shazeer, N. Generating wikipedia by summarizing long sequences. *arXiv* **2018**, arXiv:1801.10198.
22. Chen, W.; Chan, H.P.; Li, P.; Bing, L.; King, I. An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. *arXiv* **2019**, arXiv:1904.03454.
23. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P.J. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv* **2019**, arXiv:1912.08777.
24. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461.
25. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2019**, arXiv:1910.10683.
26. Harkous, H.; Groves, I.; Saffari, A. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. *arXiv* **2020**, arXiv:2004.06577.
27. Kanerva, J.; Rönnqvist, S.; Kekki, R.; Salakoski, T.; Ginter, F. Template-free data-to-text generation of Finnish sports news. *arXiv* **2019**, arXiv:1910.01863.
28. Rebuffel, C.; Soulier, L.; Scoutheeten, G.; Gallinari, P. A hierarchical model for data-to-text generation. In *European Conference on Information Retrieval*; Springer: Cham, Switzerland, 2020; pp. 65–80.
29. Kale, M. Text-to-text pre-training for data-to-text tasks. *arXiv* **2020**, arXiv:2005.10433.

30. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; Jatowt, A. YAKE! Keyword Extraction from Single Documents Using Multiple Local Features. Information Sciences. Available online: https://www.researchgate.net/publication/335766438_YAKE_Keyword_Extraction_from_Single_Documents_using_Multiple_Local_Features (accessed on 17 June 2021).
31. Gunasiri, D.Y.; Jayaratne, K.L. Automated cricket news generation in Sri Lankan style using natural language generation. *Eur. J. Comput. Sci. Inf. Technol.* **2019**, *7*, 42–56.

future internet

MDPI

Article

A Semantic Preprocessing Framework for Breaking News Detection to Support Future Drone Journalism Services

Michail Niarchos *, Marina Eirini Stamatiadou , Charalampos Dimoulas , Andreas Veglis and Andreas Symeonidis

School of Journalism & Mass Communication, Aristotle University of Thessaloniki, 54636 Thessaloniki, Greece; mstamat@jour.auth.gr (M.E.S.); babis@eng.auth.gr (C.D.); veglis@jour.auth.gr (A.V.); asymeon@eng.auth.gr (A.S.)
* Correspondence: mniarchos@jour.auth.gr

Abstract: Nowadays, news coverage implies the existence of video footage and sound, from which arises the need for fast reflexes by media organizations. Social media and mobile journalists assist in fulfilling this requirement, but quick on-site presence is not always feasible. In the past few years, Unmanned Aerial Vehicles (UAVs), and specifically drones, have evolved to accessible recreational and business tools. Drones could help journalists and news organizations capture and share breaking news stories. Media corporations and individual professionals are waiting for the appropriate flight regulation and data handling framework to enable their usage to become widespread. Drone journalism services upgrade the usage of drones in day-to-day news reporting operations, offering multiple benefits. This paper proposes a system for operating an individual drone or a set of drones, aiming to mediate real-time breaking news coverage. Apart from the definition of the system requirements and the architecture design of the whole system, the current work focuses on data retrieval and the semantics preprocessing framework that will be the basis of the final implementation. The ultimate goal of this project is to implement a whole system that will utilize data retrieved from news media organizations, social media, and mobile journalists to provide alerts, geolocation inference, and flight planning.

Keywords: breaking news; semantic processing; natural language processing (NLP); drone journalism; events location estimation

Citation: Niarchos, M.; Stamatiadou, M.E.; Dimoulas, C.; Veglis, A.; Symeonidis, A. A Semantic Preprocessing Framework for Breaking News Detection to Support Future Drone Journalism Services. *Future Internet* **2022**, *14*, 26. https://doi.org/10.3390/fi14010026

Academic Editor: Luis Javier Garcia Villalba

Received: 10 November 2021
Accepted: 6 January 2022
Published: 10 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

When it comes to event and breaking news coverage, textual representation can be supplemented with live audio–visual (AV) footage to ensure objective and reliable news reporting. While citizen and mobile journalists can offer real-time coverage through social media and User-Generated Content (UGC), they are not always capable of providing quality footage or even accessing the location of interest. On top of this, in the era of the COVID-19 pandemic, mobility is usually restricted, meaning that entering a once easily accessible place can become difficult or even impossible for a person. During the last decade, drones have democratized landscape aerial photography, making it affordable even for individual professionals and freelancers as well. Hiring or purchasing expensive equipment, such as helicopters, is no longer needed. At the same time, drones can be used to access areas that journalists would prefer not to visit while collecting high-quality AV content using their constantly improving (camera/microphone) equipment. Their relatively small size also allows their fast and effective deployment in many cases.

Moving from plain text towards AV news coverage and storytelling will improve the communication of a story [1,2]. AV data are not only valuable for the news consumers, but these data can be proven to be useful for other professionals during post-processing. This introduces the demand for high-quality footage for the data to be efficiently used by image processing tools [3–6].

Even though it is undoubted that drones can be used in several on-field scenarios, their usage is not yet commonplace. The main reason for this is that ethics and safety

concerns have led to the adoption of strict operating policies in most countries [7,8]. In addition, problems concerning intrusiveness and aesthetic ideals have been posed by photojournalists [9]. The future relaxation of the commercial application restrictions will enable the media industry and individual reporters/professionals to consistently use them. In the United States, the New York Times and the Washington Post are already using drones to take photos and videos of simulated news events at an FAA-designated test site at Virginia Tech University [10]. They have also started preparing the new generation of "dronalists", i.e., journalists trained as Pilots in Command (PICs), to use this new, exciting technology in day-to-day operations. In this spirit, this work proposes a framework for future drone journalism services, which will be fully deployable as soon as regulations have been conducted to overcome the aforementioned restrictions.

The usage of computational and robotic resources is becoming widespread in the ecosystem of Journalism 3.0. A variety of emerging journalistic forms (textual, audio, and pictorial) are enriching online media content [11]. Natural language generation (NLG) is changing production, as it can perform functions of professional journalism on a technical level [12], leading to automatic news generation [13], sometimes making use of Machine Learning (ML) and/or natural language processing (NLP) [14–17]. ML is defined as the study of algorithms that use computer systems to perform tasks without being instructed to (i.e., learning by example), while NLP, which is a part of Artificial Intelligence (AI), deals with the analysis and processing of human natural language [18]. While multi-source news retrieval is becoming commonplace, methods are in development to verify the newsgathering and the publication process [19–21]. Dedicated search and retrieval mechanisms, implemented as software agents, can be applied in data that are stored in intermediate storage locations, inferring useful information for further analysis and monitoring [22]. Even with the absence of prior knowledge, it is possible to extract events based on semantics in text and/or speech [23] or a specific timeline of events [24].

Paving the way for "dronalism", the need for the automated deployment of drones as autonomous unmanned robotic journalists has emerged. The process of news retrieval and real-time coverage of an event can be highly automated. Our hypothesis is that today's technology is mature enough to allow the implementation of a service for automated breaking news detection and notification-based communication with a drone management subsystem. The main goal is the enhancement of journalistic workflows in news coverage. More specifically, this paper aims at launching prototype automation services in drone journalism, facilitating event detection and spatiotemporal localization processes until the flight-plan execution and the news-scene coverage completion. Prototype data monitoring and recording processes will be deployed (retrieval, normalization, processing, and warehousing) to organize, classify, and annotate content/metadata properly. Stream queries and analysis will rely on the articles/posts of news-sites and social networks (blogs, Twitter, etc.), along with the transcription and indexing of press media, selected radio, and TV broadcasting programs. The aim is to accelerate human processing with machine automation that executes multiple/parallel time-, location-, and context-aware correlations in real time for event-alerting purposes. The validation of the extracted alerts and the exploitation of geographical/map information allows ground access areas to be found and recommended, where Unmanned Aerial Vehicles (UAVs) can take off towards a region of interest. Thereafter, flight-plan preparation and guided/semi-automated flight navigation and coverage can be further investigated, as long as specific criteria are met. These processes make a good fit for this current Special Issue.

One of the key elements that had to be answered from the beginning of the project is the degree to which the potential users would find the envisioned service useful, and to investigate the associated requirements and the preferred functionalities to implement, which is aligned with the analysis phase of standard software development procedures [25,26]. To achieve this, a survey was carefully designed and conducted to serve the needs of audience analysis. In this context, it is equally essential to implement and validate the applicability of the algorithmic breaking news detection back-end, which stands as the core element for

serving the whole idea. Hence, the whole concept had to be tested both in terms of users' acceptance and technologic solutions' applicability. In this direction, NLP systems were implemented as the initial algorithmic solutions and were thoroughly evaluated at various levels to provide a convincing proof of concept of the tested scenario.

The current work proposes a framework for functioning and supporting drone journalism services, incorporating the existing regulatory framework, security, and ethical matters [8,27,28]. In this context, the technological and functional capacities of drones, i.e., the capability to create powerful narratives, will be combined with a structured and automated way of operation, including alerting, control data transfer, and normalization services. The system is modular, able to integrate UGC and news alerts sent by mobile journalist devices, making use of the inherent localization and networking capabilities [25,29–33], or even cooperative drone monitoring, thus offering time-, location- and context-aware data-driven storytelling. This will lead to the faster and more reliable detection of breaking news with their contextual and spatiotemporal attributes, using crowd computer intelligence techniques with appropriate weighting mechanisms. Modularity does not only concern the retrieval and the processing stage but the output forwarding as well. Alerts will also be sent to mobile journalists to let them cover the event too. Diverse drone footage and UGC mobile streams can be combined to offer a more holistic coverage of the events.

The designed system is based on three pillars:

1. Breaking news detection and geolocation inference, which produce notifications to be sent to a drone handling system or individual drone operators.
2. Flight-plan preparation.
3. Semi-autonomous waypoint-based guidance.

Focus is given on data retrieval and, mainly in the preprocessing phase of the first stage, the whole system is built adopting a modular design approach, consisting of independent subsystems, the implementation of which is outside of this paper's focus and context.

The rest of the paper is organized as follows. A literature review around breaking news detection is presented in Section 2. Section 3 describes the assumptions made while formulating the list of actions needed, based on the opinion of field experts. The system architecture is also described in this section, along with the main functions and the components comprising the developed framework. In Section 4, we present the results of the questionnaire that was distributed for the assessment of the envisioned system, as well as the performance and the evaluation of the breaking news detection method. Conclusions are drawn and a discussion is presented in Section 5.

2. Related Work

Drones are already being used in video capturing and news coverage. Research has been carried out for flight planning and approaching inaccessible locations [34,35]. Such topics are out of the context of this work, as they refer to problems that will be faced in the last chain link of the complete system. The current stage of implementation focuses on data preprocessing and more specifically on breaking news detection; thus, the literature that is presented below refers to such semantic analysis.

One of the novelties and research contributions of the current work lies in the implemented solutions for the specific problem of multi-source breaking news detection, which makes a strong proof of concept of the proposed scenario. Based on the conducted literature/state-of-research review (and to the best of our knowledge), it seems that previous works dealing with the same problem and under similar constraints are very limited and/or entirely missing. The majority of similar approaches process social media messages (mostly tweets) [36–39], and only a few of them exploit website metadata (e.g., HTML tags [40]) or use the whole article text [41] to detect hot words. Most of the recent works seem to deal with hot topic detection, which can also be adapted to breaking news detection. However, these are two different problems in principle, i.e., "breaking news" implies urgency, while "hot topic" implies popularity. Overall, the current work proposes a holistic treatment incorporating the detection of breaking news from unstructured data emanated

by multiple sources (therefore including potential UGC contributions). The implicated spatiotemporal and contextual awareness perspectives model a semantic processing framework with modular architecture and significant future extensions on drone journalism news covering automations. A common technique for detecting hot topics is the use of keywords [36,41]. Such techniques usually show low precision, as they are based on keyword occurrence frequency in documents written by the general public [36]. The careful selection of keywords can lead to reliable results. The collection of keywords by social media posts can be very efficient. If this collection is extracted by selected reliable users, taking into consideration their influence and expertise can further improve this method [36]. Bok et al. [36] proposed the use of a modified version of TF-IDF that incorporates a temporal factor to be able to tackle the difficulty of detecting near-future breaking news.

Natural Language Inference (NLI) includes methods for predicting whether the meaning of a piece of text can be inferred by another [42]. A form of NLI is paraphrasing, which is also called text pair comparison. Computing the similarity among pieces of text is used in various applications, and it can be applied in this use case as well. Jensen–Shannon divergence is used to measure the similarity between two topics, which is used in topic tracking [43].

An interesting approach is the TDT_CC algorithm—Hot Topic Detection based on Chain of Causes [37]. This algorithm treats events as topic attributes and uses them to update a graph, aiming to detect a trend in a large graph in real time. Traditional algorithms dedicate too much time to traversing a graph, while TDT_CC tackles this issue by focusing on the structural topology.

Graphs are also used to effectively compute the relevance between textual excerpts [38]. Hoang et al. [38] utilized a term graph to measure the relevance between a tweet and a given breaking event.

Clustering is used in all of the above methods at some stage. Shukla et al. [39] applied clustering to filtered, collected tweets and then scored the clusters based on the number of tweets, which led them to breaking news detection. This is a simple and reliable technique which can only be applied on social media. In addition, this technique is not efficient in terms of memory usage and execution time, which makes it quite unusable in real-world deployments.

Applications that crawl specific websites to gather news can utilize various forms of metadata to enhance their detection methods. HTML tags constitute such a kind of metadata. They can be used to isolate the title, the subtitle, or any other semantically important piece of information to achieve a better detection rate [40]. The major defect of such an approach is that it is bound to the structure of certain platforms and is vulnerable to any possible structural change.

3. Materials and Methods

3.1. Assumptions and Setup

In order to validate our concept hypothesis regarding its user acceptance, an empirical survey was conducted in the form of a questionnaire distributed online to 100 people. Data were collected within one month (February–March 2021). Typical questions concerning breaking news capturing, reporting, and sharing were answered, retrieving vital feedback. Hence, background- and general-interest-related questions were structured in a categorical form of potential answers, with 5-point Likert scales (1–5, from "Totally Disagree" to "Totally Agree"). Binary values (i.e., gender) and higher-dimension lists were also involved. The items were divided into three subsets, with the former involving questions regarding the current state of/trends in (breaking) news reporting (Q1–Q4), the second implicating questions on the envisaged modalities and usability characteristics of the proposed service (Q5–Q9), and the latter containing basic characteristics/demographics of the users (Q10–Q17). The survey formation was validated after discussions and focus groups with representative users. Specifically, both professional and citizen journalists were involved, as well as people from the broader journalistic and news reporting field. The survey was

updated based on the received feedback, investigating the audience interest in a system that incorporates mechanisms for automatic breaking news detection and their intention to contribute with their content. The gathered information was used for the estimation of the anticipated dynamics of the proposed architectural approach. An overview of the chosen inquiries is presented here, aiming to justify the adoption and configuration of the formed questionnaire. Detailed information regarding this survey is provided in the associated Results section, along with the assessment outcomes.

During the survey preparation, all ethical approval procedures and rules suggested by the "Committee on Research Ethics and Conduct" of the Aristotle University of Thessaloniki were followed. The respective guidelines and information are available online at https://www.rc.auth.gr/ed/ (accessed on 9 July 2021). Moreover, the Declaration of Helsinki and the MDPI directions for the case of pure observatory studies were also taken into account. Specifically, the formed questionnaire was fully anonymized, and the potential participants were informed that they agreed to the stated terms upon sending their final answers. All participants had the option of quitting any time without submitting any data to the system.

3.2. Architecture

The envisioned architecture is presented in Figure 1, along with the main functions. It was designed to target the following objectives:

- To gather original data heterogeneous data sources (web, social media);
- To process and analyze input data to form an event classification taxonomy extracting breaking news;
- To create automated alerts to trigger drone applicability;
- To support semi-automated/supervised drone navigation (and tagging) through a set of waypoints (with the notes that a safe landing zone is always needed, predefined, or detected in deployment time, while PIC in the communication range is required at all times);
- To develop user-friendly dashboards for drone control;
- To synchronize aerial footage with ground-captured AV streams (e.g., UGC) and propagate related annotations based on spatiotemporal and semantic metadata (i.e., topic, timestamps, GPS, etc.);
- To investigate AV processing techniques for the detection of points/areas of interest in arbitrary images and video sequences;
- To provide information, training, and support about drone utilization in public based on local regulations and ethical codes;
- To implement semantic processing and management automation for the captured content in batch mode, which could be utilized in combination with the other publishing channels and UGC streams (i.e., for analysis, enhancement, and publishing purposes).

The architectural implementation provides a solid framework, including:

1. A prototype web service with open access to external (third-party) news sources for the detection of breaking news;
2. Pilot software automating aspects of drone-based newsgathering;
3. Content storage and management repository with indexing and retrieval capabilities;
4. On-demand support and training on the new services.

All legislation matters were accounted for during project deployment and testing and the associated training sessions. Aiming to address both breaking news detection and event geographical localization, field journalism experience was also taken into account.

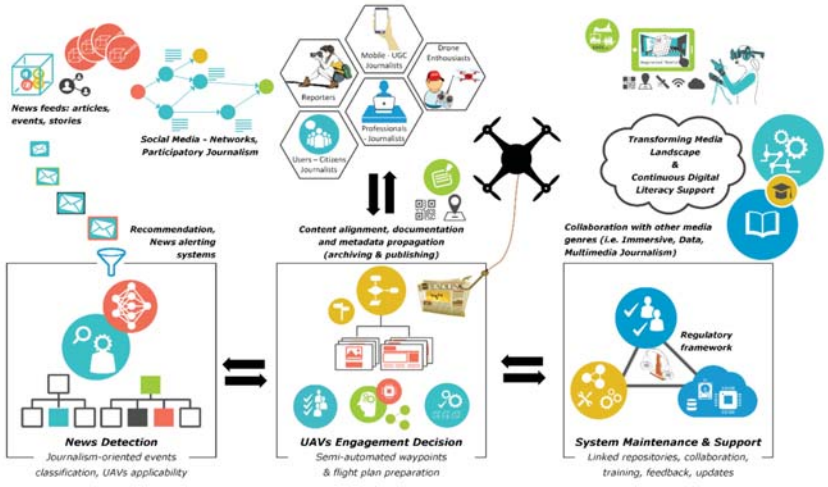


Figure 1. The envisioned concept architecture of the proposed model [8].

3.3. Framework Description

The graphical representation of the developed modular architecture is presented in Figure 2. The underlying components that drive the process of breaking news detection are described in short. Data propagation for automated, real-time drone coverage of breaking news events is presented as well. The concept was evaluated by collaborating closely with a focus group of active journalists that helped to form the outcome.

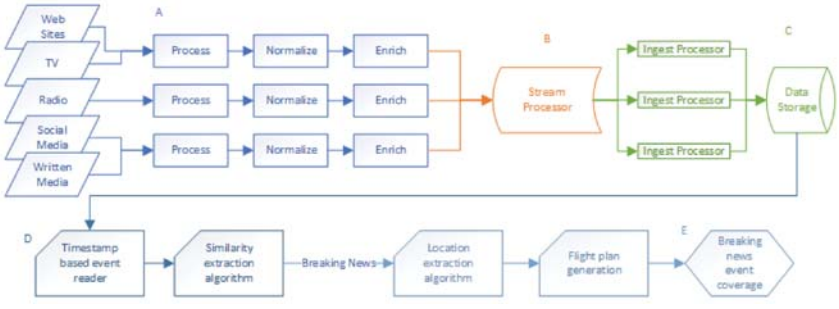


Figure 2. Block diagram of the implemented data flow.

Figure 2 depicts an overview of the implemented data flow. In the architecture, data collection starts upon their creation while monitoring several news feeds for events (A). An extensive effort was put into multi-source event gathering from (a) websites through data scraping and indexing, (b) written press such as newspapers through digitization and indexing, and (c) social media, television, and traditional radio through transcription, which is currently carried out by a news provider collaborator with the use of custom speech-to-text tools. News extracted by TV and radio is considered especially important. Many events may be firstly broadcast by these traditional media, as they still have faster access to specific news agencies. Each event is described as a single and independent data message, generated by a news feed. Such a data message, e.g., an article or a Twitter tweet, may imply one or multiple events, but in terms of news coverage, this can be considered as a single event requiring journalistic action. This initial process produces events to be propagated to the system after being enriched with additional information that allows categorization and management that will subsequently help to effectively query for events.

Events enter the system utilizing a common bus (B), a communication channel that serves the purpose of conveying information while, at the same time, ensuring data integrity and avoiding data redundancy (buffered process). It can store data in an intermediate state for a specified amount of time that can range from minutes to several days. Events that exit the common channel are stored in a database (C) in a structured way that enables fast indexing and retrieval. Events are retrieved (D) as sets of filtered information for specified time windows/ranges and checked for correlation using a multi-step detection algorithm. The output is a number of real events, along with their statistics, marked as breaking news. For a drone to be able to interconnect with such a system and operate, geographical information is needed. This information is extracted using a separate, multi-stage algorithm (E) that takes into account location-related terms in the actual text of the event, as well as related imagery that can be retrieved from it. The system is integrated as a web application that can also accept UGC that enhances user experience.

The framework consists of the modules shown in Figure 3, which outlines the conceptual architectural processes as described above. The innovation of this design lies in the ability of the system to continuously improve its performance, while also being able to handle each different input type independently by inserting new "pluggable" entities into the system (e.g., interaction with the devices of mobile journalists). Detailed information is given in the following subsections.

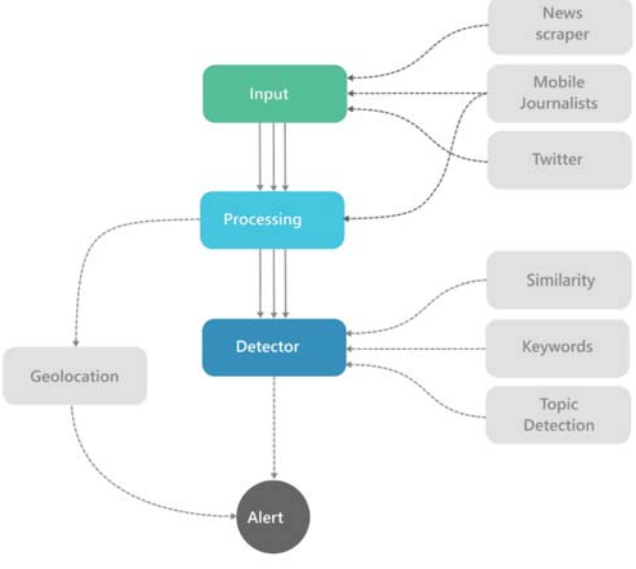


Figure 3. Modularized architecture.

Elaborating on the above, the main aim was to set up and configure a modular architecture, allowing each individual modality (radio, TV, newspaper, etc.) to be added and treated consistently, in the form of text and images. Hence, uniform text and image-derived metadata processing steps are introduced and deployed into an integrated framework. It has to be clarified that individual modules, feeding the framework with input data, are not fully optimized as part of this paper. Specifically, the back-end crawling service of a media monitoring collaborator was utilized and tested, representing a real-world scenario. As already mentioned, the associated utilities rely on OCR and speech-to-text services, supplemented with human intervention. The whole approach aligns with the objectives of the paper to provide an end-to-end pipeline encompassing seamless and uniform treatment of all individual media inputs.

3.4. Data Retrieval and Management

The process of data retrieval is critical since it provides the necessary information for the system to operate, while it introduces serious implementation difficulties. The diversity of available data sources [11] leads to the need for many different ways of handling input, while there is no lack of corner cases that require special treatment. For example, data scraping tools [44] can extract information from websites, but every website has a different structure, which means that each case requires different handling. Moreover, websites may change their presentation layer on a relatively regular basis, trying to make it more appealing to the users. Scraping tools are sensitive to changes at the presentation layer, which leads to the re-engineering of the data collection strategy. The density and the type of information included in the different data sources make the task of classification difficult. Text transcript from a radio broadcast contains less information than a simple post from a social media user profile, yet they need to be stored identically to facilitate processing capability by analytics tools. Information decomposition to title, text, source, etc., is currently executed by the news provider we are collaborating with, but ultimately, a custom implementation will be carried out to support the data retrieval from more sources.

Events are organized as sets of fields, with each one containing specific data types and content. An event schema is presented in Figure 4, along with an example event (translated in English). This process is called normalization and is the cornerstone for the creation of a unified mechanism to execute analytics [45] against really large amounts of information. The final phase taking place is the enrichment process, which can be described as the action of adding artificial fields in the original event that would not normally exist. Internal content and articles' metadata form useful information that feeds the implemented breaking news detection system. This data may include title and relevant keywords, media-related information, and time-, location-, and context-aware metadata, which in terms of NLP research community terminology are transformed to more solid and crisp representations, i.e., action, agents, place, time, etc. Despite the requirements that algorithms introduce, the events are uniformly presented in this form to also be able to propagate the input outside of this internal ecosystem. A crucial criterion behind the schema design is for it to be comprehensible by journalists that will need to process the accompanying information upon breaking news detection.

Column	Description	Type
id	Event unique identifier	integer
title	Article title	string
timestamp	Publication timestamp	string
modification_date	Last modified timestamp	string
medium_id	Medium unique identifier	integer
medium_page_views	Medium page views	integer
medium_text	Medium name	string
snippet	Snippet from the article	string
text	The whole article	string
url	URL to the article	string
is_blog	Identifies if this is a blog post	boolean

(a)


```
{
  "id": 440386725,
  "title": "European Parliament ratifies agreement on
            Brexit",
  "timestamp": "2020-01-30 13:28:39",
  "modification_date": "2020-01-30 13:30:51",
  "medium_id": 1108,
  "medium_page_views": 4617285,
  "medium_text": "thesstoday.gr",
  "snippet": "European Parliament votes ... ",
  "text": "European Parliament votes today to ratify
           on the withdrawal treaty and how it will be
           carried out ... ",
  "url": "https://thesstoday.gr/%cf%84%ce%bf-%ce%b5%cf
          %85%cf%81%cf%89%cf%80%ce%b1%cf%8a%ceb%ac%cf%8
          c-%ceb%ac%ceb%cf%bc%ceb%bd%ceb%cf%ceb%b2%ceb%bf
          %cf%8d%ceb%bb%ceb%9d%ceb%bf-%ceb%b3%cf%80%ceb%b6
          ceb%ac%cf%85%cf%81%cf%8e%ceb%bd%ceb%b5%ceb%b9/",
  "is_blog": false
}
```


(b)

Figure 4. (**a**) Event schema; (**b**) example event (json).

The output of the previous stage is a set of feeds that generate normalized events. A dedicated stream processor [6,46,47] then handles these data in a parallel manner, ensuring the integrity of the information included. By definition, a stream processor is a queueing system that concurrently accepts input from many sources (publishers) and conveys the events to processes that make use of them (subscribers/consumers). In between, a stream processor implements a buffer that can reliably store events (in transient status) for a specified amount of time (up to several weeks, provided that the required storage resources suffice). It also ensures that events are processed in the same order they entered the system. Under certain conditions, it can also ensure that the events may be replayed in the output at most, at least or exactly once, providing the required integrity. Stream processors can be set up in cluster mode, behaving like a swarm of computational resources that can replicate the information on their buffers in several different physical machines. This behavior

helps to avoid information redundancy usually existing in the buffer, making it possible to recover from software or even hardware breakdowns. Finally, it enables data retrieval by consumers that greatly improves performance through parallelization.

Events come out of the stream processor by a set of consumer processes (ingestors) that are responsible for storing the events in a long-term logbook. The used data format facilitates the storage of massive amounts of data in cloud providers while improving retrieval by sharing data files and optimizing filtered read queries. TileDB [48,49] was chosen as the database system that is responsible for the above procedure. This data engine treats input events as elements in a three-dimensional array, defined as follows: The first dimension is time, expressed in seconds since a specific date and time in the past, mapping to the timestamp field of the event. The second dimension is the actual medium (source) of events. There exists a predefined list of sources identified by special identifiers in integer format. The third and last dimension is the event identifier, which is unique for a specific event and helps in discriminating events that have the same timestamp (measured in seconds). The rest of the available fields of an event are stored in a dedicated cell with coordinates of timestamp, medium, and event as required by the dimensions of the array as attributes. Having the events securely stored using an efficient format allows for their usage in the context of a breaking news reasoner/extractor.

3.5. Breaking News Discovery

Following the setup that was created in Section 3.1, the next step is the utilization of available data that are stored in the medium of our choice. The target is to infer breaking news from the total list of available events. It was decided that a reliable approach would consist of both NLP techniques and expert knowledge; thus, the designed component combines standard computational methods with empirical rules.

Long-time journalists and reporters were engaged to formulate the breaking news detection procedure. Table 1 provides a questionnaire that was initially formed in technical terms (i.e., concerning system implementation) and was further elaborated by incorporating the feedback given by experienced newsroom professionals. While the utmost target is to address these queries to a significant number of media professionals and citizen journalists, an initial validation was obtained by ten (10) targeted users/professionals, different from those involved in the initial formulation process.

The intention was to proceed with an initial parameterization and setup of the proposed system that could be further configured and fine-tuned, upon receiving more audience feedback and testing experience.

The prototype algorithm that was implemented is presented in Algorithm 1. The core of the algorithm is the measurement of the similarity [50–52] of the event title under test with the rest of the titles that have appeared in the near past.

The main advantage of this technique is that it does not require any prior knowledge of the processed text, as it uses a model that has already been trained on a set of documents that serves as a knowledge base. A big data set (500k words) in the Greek language is obtained and then projected to an applicable vector space [53]. Each word in the corpus is represented by a 300-dimensional vector. The vectors are calculated using the standard GloVe algorithm [54], which is a global log-bilinear regression model that combines global matrix factorization and local context window methods. This algorithm is widely used as it performs quite well on similarity tasks. The extracted vectors that are also called word embeddings contain not only the meaning of the word as an arithmetic value but its relation to other words as well. The described process is unsupervised, capable of learning by just analyzing large amounts of texts without user intervention. Moreover, it is possible to use transfer learning—a practice in machine learning where stored knowledge gained while solving one problem can be applied to a different but related problem—on an existing trained data set using a lexicon of general interest to extend its functionality, to achieve a better-contextualized word representation in specific scenarios, where different buzzwords/lexicons are used (e.g., a journalistic dictionary).

Table 1. Empirical rules regarding the criteria for announcing an event as emerging.

Survey Question	Empirical Results, Subject to Test
In your experience, how many common words should the title of two articles contain to be considered to refer to the same event?	One-four or over five words, depending on the length of the title.
In your experience, what is the maximum time difference of the articles under comparison to being considered as referring to the same extraordinary event?	From thirty minutes to over three hours, with the shortest duration to provide stronger indications. However, this does not matter if it is an extraordinary event it will be constantly updated.
When developing an algorithmic system for automated comparison between different articles, would it make sense to use?	Multiple sequential or sliding time-windows, with a degree of overlapping.
What criteria would you use to determine the importance of an article?	Number of sources that simultaneously appear, number of "reliable" sources, specific thematic classification (e.g., natural disaster), number of sharing posts and/or reactions (likes, comments, etc.).
What extra fields would you consider important for the purpose?	Author's name listed, images attached/enclosed, sources and their reliability.

Each title is also represented by a vector, which is calculated by averaging the vectors of the words that constitute it. To measure the relevance between two titles, the cosine similarity between the corresponding vectors is computed [36,39,41,42].

$$S_c(A, B) = \frac{A \times B}{||A||\ ||B||} \quad (1)$$

Cosine similarity of zero means that two given sentences have no relevance, while a value of one implies an absolute match. Various experiments were conducted to define the minimum value that implies high contextual relevance. The results of these experiments were given to the field experts that were mentioned earlier, and they set the minimum threshold to the value of 0.85.

Algorithm 1 Breaking news detection

```
1: while forever do
2:   ▷ get the articles of the past h hours
3:   get articles["timestamp >= now-h"]
4:   for article in articles do
5:     int counter
6:     title ← get_title(article)
7:     title_vector ← get_vec(title)
8:     for tmp_article in articles do
9:       tmp_title ← get_title(tmp_article)
10:          tmp_title_vector ← get_vec(tmp_title)
11:          s ← similarity(title_vector, tmp_title_article)
12:          if s >= t then
13:            increment counter
14:          end if
15:       end for
16:       if counter >= m then
17:          classify article as breaking
18:       end if
19:     end for
20: end while
```

The detection algorithm is now at the proof of concept stage while it continues to be improved. We are aiming to introduce more modules that will be part of the "detector" super module, each one contributing to the process differently. A topic detection model will also be integrated to improve the filtering of the "candidate" events. The problem of topic detection is not solved, but there have been promising works that introduced automated classification methods [55] that can fit in the proposed framework. On top of this, given that classification algorithms require continuous testing and training using data from reliable databases, focus is given to the continuous enrichment of the designed database, which will help in improving the detection methods.

3.6. Geolocation Inference

The next step is to decide whether the event is eligible for real-time drone coverage or even if a flight plan can be initially designed using waypoints. An important module that will be plugged into the system is the geolocation module. If input data do not contain any GPS tags, then the location has to be algorithmically extracted from the article without compromising the required automation. Combining the knowledge extracted by the group of the ten (10) experts, mentioned in the previous section, and regarding methods mentioned in the recent literature led to the formation of the following list of methods:

- Geo-tags from the associated sources [56];
- Geolocation extraction from named entities, such as [19]:
    - Geopolitical entities (i.e., countries, cities, and states);
    - Locations (i.e., mountains and bodies of water);
    - Faculties (i.e., buildings, airports, highways, etc.);
    - Organizations (i.e., companies, agencies, institutions, etc.).
- Photo documents of the article, through inverse image search/near-duplicate image retrieval [57], followed by visual semantics recognition [58] and geographical metadata propagation.

The above approaches can work in both a fully and semi-automated fashion, expediting the searching process, which can be substantially propelled by human cognition and experience. The incorporation of mobile users (UGC contributions) could assist in the targeted location awareness. In all cases, the system provides an output that optionally includes location information or alerts the user if it is not possible to extract such information.

At the moment, a related base algorithm has been designed and put into a test, making an initial proof of concept. Preliminary results have been extracted, but they still cannot be presented, as the reliability of the implementation is still under dispute.

3.7. Topic Detection

An additional layer of "filtering" is the topic detection/classification module. Topics of specific categories should never appear as breaking news, e.g., lifestyle news. For this reason, NLP and ML methods [59] have been implemented to achieve reliable topic classification and to reduce the input size that will be fed to the detector module. Apart from filtering, this makes it possible to apply weighting based on the topic, as some topics are less likely to be breaking.

In our current approach, the vector space models of the documents are calculated as soon as the documents have been preprocessed (stop-word elimination, stemming, etc.). After this step, clustering is applied based on the cosine similarity of the vectors [60]. This approach may sound simplistic but gives good results when there are only a few topics. The involvement of field experts in the implementation procedure enables the design of a rule-based detection system [42]. Such an approach can be easily implemented from a programming point of view, but it requires deep knowledge of the domain, which makes maintainability difficult. This implementation is still in the tuning phase, as the first conducted evaluation has shown that the percentage of unclassified articles is quite high.

Eventually, deep learning approaches [37] are going to be implemented and compared with the aforementioned one to further evaluate its performance. The main advantage of such methods is that they do not require deep domain knowledge, which leads to faster implantation and makes maintenance and further development much easier. Beyond traditional deep learning techniques, there are novel methods such as Deep Reinforced Learning [21], which can be utilized to achieve even better results.

4. Experimental Results

4.1. Concept Validation through Audience Analysis

To examine audience acceptance of the proposed approach, we undertook an online survey (N = 100). Online surveys allow for generalized data collection, and they are proven to be feasible methods to reach a broad and representative sample. Table 2 synopsizes the final set of questions selected for the needs of this survey. In general, the results showed that many people are not familiar with what DJ is. The majority of the participants expressed their interest in the mediated breaking news experience that is aimed to be achieved within the current project, as thoroughly analyzed below.

Table 2. The analysis questionnaire was answered by 100 people. (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree).

	Questions Answered in Likert Scale	1	2	3	4	5
Q1	I am interested in breaking news	0	4%	15%	40%	41%
Q2	Cross check upon breaking news alert is necessary	0	0	8%	46%	46%
Q3	An automated breaking news alert system would be useful	1%	6%	16%	59%	18%
Q4	I would use an automated breaking news alert system	2%	4%	19%	53%	22%
Q5	I am familiar with the term "drone journalism"	46%	21%	18%	9%	6%
Q6	Citizen journalists' contribution with multimodal data (image, video, text) makes breaking news detection easier	2%	4%	24%	47%	23%
Q7	Contribution with multimodal data (image, video, text) by citizen journalists' that use mobile devices with geo-tagging capabilities makes breaking news detection easier	0	4%	16%	60%	20%
Q8	Citizen journalists' contribution with multimodal data (image, video, text) makes local breaking news reporting easier	0	2%	23%	60%	15%
Q9	Citizen journalists' contribution with multimodal data (image, video, text) leads to misconceptions	2%	15%	41%	31%	11%

In total, 60% of the respondents are not familiar with the DJ term. However, 85% of them would use an automated breaking news reporting system similar to the one presented, while 83,3% of them believe that such a system would be useful. The system is further validated by the fact that 77% believe that such a system would be useful, while 77% would use it for automated breaking news reporting. In total, 55% believe that crowdsourced data coming from citizen journalists that use mobile devices would offer better news coverage, and the same percentage (55%) believe that GPS-enabled devices would offer better news localization. This primitive user acceptance is also evaluated by the fact that 58% of the respondents do not see crowdsourced multimodal content as a threat but rather as an opportunity for better data management of breaking news.

Demographics

Out of 100 respondents, 27% are male and 73% are female. As far as the age groups to which the respondents belong, 58%, 31%, and 11% of them belong in the age group of 18–25, 26–40, and above 40, respectively. Most of the respondents' roles (74%) are that of news consumers, while only 9% have a journalistic role (either as professional journalists or professionals in the broader mass communication field). In sum, 17% of the collected responses come from people that consider themselves as dynamic news consumers and/or distributors (via personal pages or blogs). Almost all of the respondents (99%) use the

internet at least once a day, and 78% of them use the internet for news retrieval. Figure 5 below, also presents the MEAN and Standard Deviation values for questions Q1 and Q5 – Q9, according to the results discussed above.

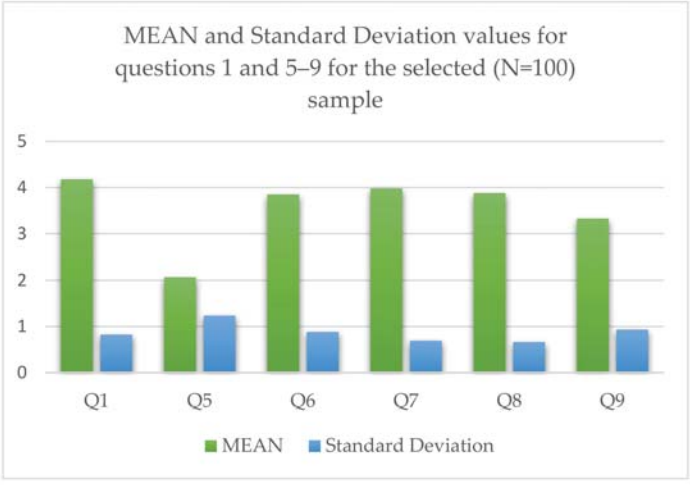


Figure 5. Graph statistics for the respondents' group (N = 100), concerning some of the important questions (namely, Q1 and Q5–Q9). It is noteworthy that although DJ is not that widespread, the participants showed a high interest in the proposed application. This can be further supported by the fact that the answers' MEAN values to the questions Q6–Q8 is near 4 ("Agree"), with S.D. values kept below 1.

In summary, the results of the conducted survey validate the research hypothesis that there is an audience willing to share (own) crowdsourced content, contributing to the easier data management of breaking news, that will provide enhanced DJ interfaces.

4.2. Qualitative System Validation Using Experts (News and Media Professionals)

While the implementation of the major components of the system was completed, the presented framework is a work in progress concerning its overall integration. Data gathering, processing, normalization, propagation to the pipeline, and storage are already in place. Moreover, similarity results are being tested end-to-end, while the algorithm to extract breaking news is under evaluation using a full load of input events. Geo-tag extraction is underway along with a flight plan and control system based on a web user interface. The aforementioned modalities and the system as a whole were validated through formative evaluation sessions, interviewing experts in the field, working in newsrooms, and manually performing most of the underlying tasks. The results are presented in Table 3 and Figure 6, in which the mean and standard deviation values of all answers are calculated. Based on these preliminary results, the usefulness of a system that automatically compares articles for the purpose of breaking news detection can be evaluated and conclusions can be extracted, utilities that most of the sample evaluators/experts would use. Lower results in the trust column are considered expected since the system is under construction and has to prove its applicability and accuracy in practice, even though there is high confidence regarding its performance in the field. It is expected that the gradual elaboration of the systems and the associated datasets would further increase the robustness and efficiency of the service in different real-life scenarios.

Table 3. Questionnaire answered, evaluating the acceptance of presented work by experts (E1–E10) (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree).

	Question	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
EQ1	Would you be interested in a system that collects all the news articles to create a data set?	4	5	4	3	3	4	5	5	4	5
EQ2	Would you be interested in a system to look up information?	2	3	2	2	3	3	3	4	4	5
EQ3	Would you be interested in a system that would allow you to compare articles with each other?	5	5	4	4	4	5	4	5	5	5
EQ4	Would you be interested in an automatic emergency detection system?	3	5	5	4	4	3	4	4	3	4
EQ5	An automatic breaking news system is useful.	3	4	4	4	4	4	4	4	3	4
EQ6	I would use an automatic emergency detection system.	3	5	5	4	4	4	4	4	4	5
EQ7	I would trust an automatic emergency detection system.	2	3	3	3	3	3	3	4	4	4

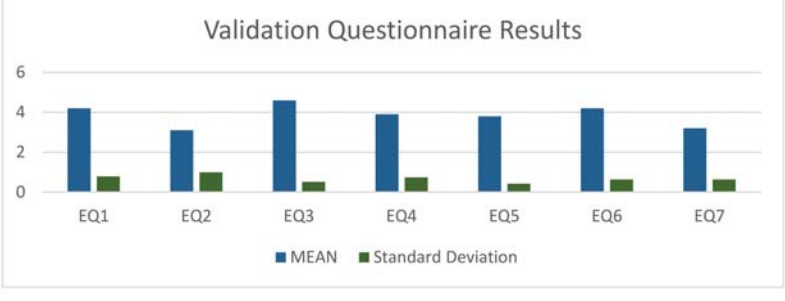


Figure 6. Mean and standard deviation of the answers given in the validation questionnaire.

4.3. Preliminary System Evaluation on Real-World Data

During the evaluation process, the need for a news database was raised. Due to the inexistence of a public and open Greek media news database, a custom one was constructed and populated, consisting of 38,759 articles, gathered in 175 days from 412 online news platforms. A custom tool for news-feed collection from real-life aggregators was developed, aiming to provide the deployed database with necessary data.

Each article was classified as "breaking" or "not breaking" by assigning the respective label to it. Classification could only be carried out by humans; thus, three experts were invited to process the article titles of two whole dates (526 articles) and place a "vote" for each one they perceived as breaking [61]. Two testing sets were created, namely *three-vote unanimous* and the *two-vote majority*. The breaking events of the former set were voted by all three experts as such, whilst the breaking events of the latter were voted by at least two experts. The dataset is publicly available on http://m3c.web.auth.gr/research/datasets/bndb/ (accessed on 9 November 2021).

After constructing this first testing dataset and running the detection tool on it, we determined the evaluation metrics that should be used. Precision and recall are two of the most commonly used metrics in information retrieval and classification [62]. Precision is the fraction of the documents retrieved that are relevant to the user's information need:

$$\text{precision} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (2)$$

Recall is the fraction of the documents that are relevant to the query that is successfully retrieved:

$$\text{recall} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (3)$$

The Precision–Recall curve is one of the most used metrics for the evaluation of binary classifiers. In the current use-case, this curve is affected by the involved system parameters, namely time window (*tw*) that controls observation length, similarity index threshold (*tsi*) that estimates the relation of the retrieved articles, and count threshold (*cnt*) that summates the number of the retrieved documents within the wanted *tsi* values. This was initially checked through empirical observations, with trial-and-error testing to indicate an appropriate time window of *tw* = 1800 s (i.e., half an hour). Hence, it was decided to further elaborate on the behavior of the classifier through the Precision–Recall curve concerning the change of the similarity threshold (Figure 7) and the count threshold (Figure 8). It is expected that greater threshold values will have a great impact on Precision, as the detection algorithm will be "stricter".

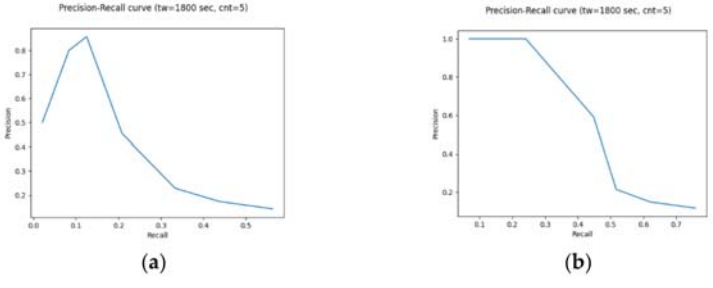


Figure 7. Precision–Recall curve keeping the window at 1800 s and the count threshold at 5 while increasing the similarity threshold from 0.55 to 1.0 with a step of 0.5: (**a**) three-vote unanimous set; (**b**) two-vote majority set.

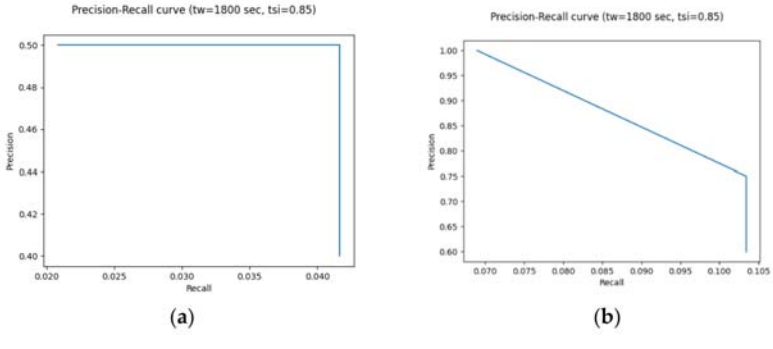


Figure 8. Precision–Recall curve keeping the window at 1800 s and the similarity threshold at 0.85, while increasing the count threshold from 3 to 13 with a step of 1: (**a**) three-vote unanimous set; (**b**) two-vote majority set.

An attempt to evaluate the behavior of the detection system was also made by changing the time window from 600 to 6000 s while keeping the similarity threshold at 0.85 and the count threshold at 5, but the results remained the same.

Overall, based on the evaluation results of the trained models, the parameter that has the greatest impact on the classifier's behavior is the similarity threshold. Setting the similarity threshold to 0.75, a Precision value of 0.857 and a Recall value of 0.125 were achieved on the three-vote unanimous set. This leads to the conclusion that most of the retrieved articles were truly classified as breaking news, whilst the probability of a breaking article to be retrieved by the system was kept low. This prototype system performs remarkably better on the two-vote majority set, on which setting the similarity threshold to 0.75 a precision value of 1.0, and a recall value of 0.241 is achieved.

Recalling the conducted state-of-research review, related works, which deal either with hot topic detection or breaking news detection, evaluate their results with the same metrics, achieving average Precision and Recall values of 0.84 and 0.83, respectively [36,40]. It should be highlighted that the accuracy of the systems that try to solve these problems is still relatively low at the level of 80% [36,40,41]. Given that the current work faces a broader/more demanding problem, and the proposed solution fertilizes the ground for future ML/DL elaborations (the full potential of which have not been explored at this point), we can support the idea that the results achieved are mostly adequate, strongly validating the targeted proof of concept.

5. Discussion

The aforementioned preliminary results of the implemented prototype were also given to the field experts mentioned above. Based on their feedback, they are promising, but quite a few improvements are required. As the database will be growing bigger and more modules will be implemented and integrated, the results will get better. A bigger database will not only offer the opportunity of moving to deep learning methods but will also assist in the introduction of media source weights. Some media are more reliable than others, which is something that should be taken into consideration by the implemented logic. It also seems that a custom corpus should be built for training the word vector providing model, as words in the news continuously change, and a static lexicon will not be able to cover future cases.

The integration of more input sources will also improve the results, as social media will provide a greater or at least a different kind of insight than that of the websites. End users should also be able to choose specific social media accounts to monitor so as not to burden the system with a high volume of useless data.

The human factor will play a very serious role in the efficiency of this framework. Mobile journalists that will be trusted with both sending alerts to and receiving alerts from the platform will be carefully chosen. Reliability is the most important trait for both functions, which means that the selection should be made with the assistance of professionals.

6. Future Work

As soon as the breaking news detection method reaches an ideal state, the next step will be to work on the other two pillars. Flight-plan preparation demands robust and reliable geolocation inference. The next milestone will be to provide geo-location information extracted by the content of the event entities in case they are not explicitly included in them. It may be impossible to infer the exact location of interest just from textual or visual data, but humans will always be involved in the process to correct or improve it. This stage includes the trajectory design as well as logistics operations, such as drone and operator availability checking as well as requesting flight authorization if required by the law.

The last step will be the provision of semi-autonomous guiding to the drone or the human operator through waypoints. This process will demand access to three-dimensional maps and advanced image processing for obstacle avoidance.

7. Conclusions

Drones are very useful in hostile or hazardous environments where the human approach is considered dangerous and unsafe or when ground coverage is not feasible; there is also the “disposable drone scenario”, where it is expected that UAVs might not return to the land zone. Besides the aims of healthier and less dangerous working conditions for reporters, new storytelling methods can be exploited to raise public awareness on specific news stories (e.g., a conflict in a war zone, etc.) Extending the above, drone journalism could be the ideal tool for reporting physical disasters while delivering valuable civil protection informing services to the public (e.g., monitoring an earthquake or hurricane, traffic jam, etc.) Additional cases include environmental journalism, nature inspection, wildlife protection, and sports coverage.

Future drone journalism services will be able to address numerous challenges that people are already facing. For instance, environmental issues and the better management of earth resources are considered very critical for the upcoming decades. Journalists and news media are obliged to inform people and to help them realize the actual situation as well as actions that should be taken. UAV surveillance allows for long-term environmental observation, which could be quite useful in the above directions. New technologies (i.e., multispectral imaging) will allow the monitoring of invisible changes, which can be conducted regularly. While the technology of multi-spectral vision is currently expensive, it is more than certain that such imaging tools will be commonly available in the not-so-far future.

The proposed framework is in alignment with important worldwide initiatives, dedicated to establishing the ethical, educational and technological framework for this emerging field. People deserve to be properly informed and to be given further assistance on arising social, techno-economic, legal, and moral matters. Another positive impact is the implementation of smart systems and the acquisition of large-scale annotated content repositories, featuring common ground with far-reaching aims, such as the Semantic/Intelligent Web (Web 3.0/4.0) and the Internet of Things.

The benefits of such a system may be numerous, but amateur and professional users must be careful when using devices that can record video and audio. Personal data should be respected by everyone and under all circumstances. Every day, people come before cameras that are intentionally or unintentionally pointing at them. Drones have a longer range, which means that footage that includes people should be automatically censored before being publicly broadcasted. Every action should be governed by journalistic ethics.

Author Contributions: Conceptualization, M.N. and C.D.; methodology, M.N. and M.E.S.; software, M.N.; validation, M.N. and C.D.; formal analysis, M.N., C.D and A.S.; investigation, M.E.S.; resources, C.D., A.S. and A.V.; data curation, C.D., A.S. and A.V.; writing—original draft preparation, M.N. and M.E.S.; writing—review and editing, C.D., A.S. and A.V.; visualization, M.N. and M.E.S.; supervision, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data that are not subjected to institutional restrictions are available through the links provided within the manuscript.

Acknowledgments: The authors acknowledge the valuable contribution of Innews S.A. (https://www.innews.gr/, accessed on 9 November 2021) for providing us access to their repository with original indexed source data/articles that helped in proving the initial hypothesis. They also acknowledge Software Engineer Andreas Ntalakas for envisioning and setting up the system architecture design.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Drones and Journalism: How Drones Have Changed News Gathering. Available online: https://www.simulyze.com/blog/drones-and-journalism-how-drones-have-changed-news-gathering (accessed on 23 March 2017).
2. Taking Visual Journalism into the Sky with Drones. Available online: https://www.nytimes.com/2018/05/02/technology/personaltech/visual-journalism-drones.html (accessed on 2 May 2018).
3. Gynnild, A. The robot eye witness: Extending visual journalism through drone surveillance. *Digit. J.* **2014**, *2*, 334–343. [CrossRef]
4. Hirst, M. *Navigating Social Journalism: A Handbook for Media Literacy and Citizen Journalism*, 1st ed.; Routledge: Abingdon, UK, 2019.
5. How Drones Can Influence the Future of Journalism. Available online: https://medium.com/journalism-innovation/how-drones-can-influence-the-future-of-journalism-1cb89f736e86 (accessed on 17 December 2016).
6. Palino, T.; Shapira, G.; Narkhede, N. *Kakfa: The Definitive Guide*; O'Reilly: Newton, MA, USA, 2017.
7. Dörr, K.N.; Hollnbuchner, K. Ethical Challenges of Algorithmic Journalism. *Digit. J.* **2017**, *5*, 404–419. [CrossRef]
8. Ntalakas, A.; Dimoulas, C.A.; Kalliris, G.; Veglis, A. Drone journalism: Generating immersive experiences. *J. Media Crit.* **2017**, *3*, 187–199. [CrossRef]

9. Harvard, J. Post-Hype Uses of Drones in News Reporting: Revealing the Site and Presenting Scope. *Media Commun.* **2020**, *8*, 85–92. [CrossRef]
10. Virginia Tech. Mid-Atlantic Aviation Partnership. Available online: https://maap.ictas.vt.edu (accessed on 13 December 2018).
11. Valchanov, I.; Nikolova, M.; Tsankova, S.; Ossikovski, M.; Angova, S. *Mapping Digital Media Content. New Media Narrative Creation Practices*; University of National and World Economy: Sofia, Bulgaria, 2019.
12. Dörr, K.N. Mapping the field of Algorithmic Journalism. *Digit. J.* **2015**, *4*, 700–722. [CrossRef]
13. Haim, M.; Graefe, A. Automated news: Better than expected? *Digit. J.* **2017**, *5*, 1044–1059. [CrossRef]
14. Fillipidis, P.M.; Dimoulas, C.; Bratsas, C.; Veglis, A. A unified semantic sports concepts classification as a key device for multidimensional sports analysis. In Proceedings of the 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Zaragoza, Spain, 6–7 September 2018; pp. 107–112.
15. Fillipidis, P.M.; Dimoulas, C.; Bratsas, C.; Veglis, A. A multimodal semantic model for event identification on sports media content. *J. Media Crit.* **2018**, *4*, 295–306.
16. Shangyuan, W.; Edson, C.T., Jr.; Charles, T.S. Journalism Reconfigured. *J. Stud.* **2019**, *20*, 1440–1457. [CrossRef]
17. Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. [CrossRef]
18. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. 2014. Available online: https://nlp.stanford.edu/pubs/glove.pdf (accessed on 9 November 2021).
19. Katsaounidou, A.; Dimoulas, C. Integrating Content Authentication Support in Media Services. In *Encyclopedia of Information Science and Technology*, 4th ed.; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA, 2017.
20. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018.
21. Shahbazi, Z.; Byun, Y.C. Fake Media Detection Based on Natural Language Processing and Blockchain Approaches. *IEEE Access* **2021**, *9*, 128442–128453. [CrossRef]
22. Symeonidis, A.L.; Mitkas, P.A. *Agent Intelligence through Data Mining*; Springer Science & Business Media: Berlin, Germany, 2006.
23. Xiang, W.; Wang, B. A Survey of Event Extraction from Text. *IEEE Access* **2019**, *7*, 173111–173137. [CrossRef]
24. Piskorski, J.; Zavarella, V.; Atkinson, M.; Verile, M. Timelines: Entity-centric Event Extraction from Online News. In Proceedings of the Text 2 Story 20 Workshop 2020, Lisbon, Portugal, 14 April 2020; pp. 105–114.
25. Stamatiadou, M.E.; Thoidis, I.; Vryzas, N.; Vrysis, L.; Dimoulas, C. Semantic Crowdsourcing of Soundscapes Heritage: A Mojo Model for Data-Driven Storytelling. *Sustainability* **2021**, *13*, 2714. [CrossRef]
26. Chatzara, E.; Kotsakis, R.; Tsipas, N.; Vrysis, L.; Dimoulas, C. Machine-Assisted Learning in Highly-Interdisciplinary Media Fields: A Multimedia Guide on Modern Art. *Educ. Sci.* **2019**, *9*, 198. [CrossRef]
27. Drone Journalism: Newsgathering Applications of Unmanned Aerial Vehicles (UAVs) in Covering Conflict, Civil Unrest and Disaster. Available online: https://assets.documentcloud.org/documents/1034066/final-drone-journalism-during-conflict-civil.pdf (accessed on 10 October 2016).
28. Culver, K.B. From Battlefield to Newsroom: Ethical Implications of Drone Technology in Journalism. *J. Mass Media Ethics* **2014**, *29*, 52–64. [CrossRef]
29. Sidiropoulos, E.A.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C.A. Collecting and Delivering Multimedia Content during Crisis. In Proceedings of the EJTA Teacher's Conference 2018, Thessaloniki, Greece, 18–19 October 2018.
30. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C.A. A mobile cloud computing collaborative model for the support of on-site content capturing and publishing. *J. Media Crit.* **2018**, *4*, 349–364.
31. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. jReporter: A Smart Voice-Recording Mobile Application. In Proceedings of the 146th Audio Engineering Society Convention, Dublin, Ireland, 20–23 March 2019.
32. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How in Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173. [CrossRef]
33. Vryzas, N.; Sidiropoulos, E.; Vrysis, L.; Avraam, E.; Dimoulas, C. Machine-assisted reporting in the era of Mobile Journalism: The MOJO-mate platform. *Strategy Dev. Rev.* **2019**, *9*, 22–43. [CrossRef]
34. Petráček, P.; Krátký, V.; Saska, M. Dronument: System for Reliable Deployment of Micro Aerial Vehicles in Dark Areas of Large Historical Monuments. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2078–2085. [CrossRef]
35. Krátký, V.; Alcántara, A.; Capitán, J.; Štěpán, P.; Saska, M.; Ollero, A. Autonomous Aerial Filming With Distributed Lighting by a Team of Unmanned Aerial Vehicles. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7580–7587. [CrossRef]
36. Bok, K.; Noh, Y.; Lim, J.; Yoo, J. Hot topic prediction considering influence and expertise in social media. *Electron. Commer Res.* **2019**, *21*, 671–687. [CrossRef]
37. Liu, Z.; Hu, G.; Zhou, T.; Wang, L. TDT_CC: A Hot Topic Detection and Tracking Algorithm Based on Chain of Causes. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*; Springer: Cham, Switzerland, 2018; Volume 109, pp. 27–34. [CrossRef]
38. Hoang, T.; Nguyen, T.; Nejdl, W. Efficient Tracking of Breaking News in Twitter. In Proceedings of the 10th ACM Conference on Web Science (WebSci'19), New York, NY, USA, 26 June 2019; pp. 135–136. [CrossRef]
39. Shukla, A.; Aggarwal, D.; Keskar, R. A Methodology to Detect and Track Breaking News on Twitter. In Proceedings of the Ninth Annual ACM India Conference, Gandhinagar, India, 21–23 October 2016; pp. 133–136. [CrossRef]

40. Jishan, S.; Rahman, H. Breaking news detection from the web documents through text mining and seasonality. *Int. J. Knowl. Web Intell.* **2016**, *5*, 190–207. [CrossRef]
41. Zhu, Z.; Liang, J.; Li, D.; Yu, H.; Liu, G. Hot Topic Detection Based on a Refined TF-IDF Algorithm. *IEEE Access* **2019**, *7*, 26996–27007. [CrossRef]
42. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv* **2021**, arXiv:2004.03705. [CrossRef]
43. Xu, G.; Meng, Y.; Chen, Z.; Qiu, X.; Wang, C.; Yao, H. Research on Topic Detection and Tracking for Online News Texts. *IEEE Access* **2019**, *7*, 58407–58418. [CrossRef]
44. Web Scrapping Using Python and Beautiful Soup. Available online: https://towardsdatascience.com/web-scraping-5649074f3ead (accessed on 20 March 2020).
45. Avraam, E.; Veglis, A.; Dimoulas, C. Publishing Patterns in Greek Media Websites. *Soc. Sci.* **2021**, *10*, 59. [CrossRef]
46. Dean, A.; Crettaz, V. *Event Streams in Action*, 1st ed.; Manning: Shelter Island, NY, USA, 2019.
47. Psaltis, A. *Streaming Data*, 1st ed.; Manning: Shelter Island, New York, NY, USA, 2017.
48. Papadopoulos, S.; Datta, K.; Madden, S.; Mattson, T. The TileDB array data storage manager. *Proc. VLDB Endow.* **2016**, *10*, 349–360. [CrossRef]
49. TileDB. Available online: https://docs.tiledb.com/main/ (accessed on 23 January 2021).
50. Guo, W.; Zeng, Q.; Duan, H.; Ni, W.; Liu, C. Process-extraction-based text similarity measure for emergency response plans. *Expert Syst. Appl.* **2021**, *183*, 115301. [CrossRef]
51. Yang, S.; Huang, G.; Ofoghi, B.; Yearwood, J. Short text similarity measurement using context-aware weighted biterms. *Concurr. Comput. Pract. Exp.* **2020**, e5765. [CrossRef]
52. Shahmirzadi, O.; Lugowski, A.; Younge, K. Text Similarity in Vector Space Models: A Comparative Study. In Proceedings of the 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 17 February 2020; pp. 659–666. [CrossRef]
53. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
54. Azunre, P. *Transfer Learning*; Manning: Shelter Island, NY, USA, 2021.
55. Bodrunova, S.S.; Orekhov, A.V.; Blekanov, I.S.; Lyudkevich, N.S.; Tarasov, N.A. Topic Detection Based on Sentence Embeddings and Agglomerative Clustering with Markov Moment. *Future Internet* **2020**, *12*, 144. [CrossRef]
56. Middleton, S.E.; Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, Y. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Trans. Inf. Syst.* **2018**, *36*, 40. [CrossRef]
57. Dong, W.; Wang, Z.; Charikar, M.; Li, K. High-confidence near-duplicate image detection. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 5–8 June 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 1–8.
58. Li, X.; Larson, M.; Hanjalic, A. Geo-distinctive visual element matching for location estimation of images. *IEEE Trans. Multimed.* **2017**, *20*, 1179–1194. [CrossRef]
59. Li, Z.; Shang, W.; Yan, M. News Text Classification Model Based on Topic model. In Proceedings of the IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016; p. 16263408. [CrossRef]
60. Patel, S.; Suthar, S.; Patel, S.; Patel, N.; Patel, A. Topic Detection and Tracking in News Articles. In Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, 25–26 March 2017. [CrossRef]
61. Dimoulas, C.; Papanikolaou, G.; Petridis, V. Pattern classification and audiovisual content management techniques using hybrid expert systems: A video-assisted bioacoustics application in Abdominal Sounds pattern analysis. *Expert Syst. Appl.* **2011**, *38*, 13082–13093. [CrossRef]
62. Rinaldi, A.M.; Russo, C.; Tommasino, C. A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features. *Future Internet* **2020**, *12*, 183. [CrossRef]

future internet

MDPI

Article

Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets

Andreas Giannakoulopoulos [1,*], Minas Pergantis [1,*], Nikos Konstantinou [1], Alexandros Kouretsis [1], Aristeidis Lamprogeorgos [1] and Iraklis Varlamis [2]

[1] Department of Audio and Visual Arts, Ionian University, 7 Tsirigoti Square, 49100 Corfu, Greece; nikoskon@ionio.gr (N.K.); akourets@gmail.com (A.K.); a18labr@ionio.gr (A.L.)
[2] Department of Informatics and Telematics, Harokopio University of Athens, 70 El. Venizelou Str., 17676 Athens, Greece; varlamis@hua.gr
* Correspondence: agiannak@ionio.gr (A.G.); a19perg6@ionio.gr (M.P.)

Abstract: Since the dawn of the new millennium and even earlier, a coordinated effort has been underway to expand the World Wide Web into a machine-readable web of data known as the Semantic Web. The field of art and culture has been one of the most eager to integrate with the Semantic Web, since metadata, data structures, linked-data, e.g., the Getty vocabularies project and the Europeana LOD initiative—and other building blocks of this web of data are considered essential in cataloging and disseminating art and culture-related content. However, art is a constantly evolving entity and as such it is the subject of a vast number of online media outlets and journalist blogs and websites. During the course of the present study the researchers collected information about how integrated the media outlets that diffuse art and culture-related content and news are to the Semantic Web. The study uses quantitative metrics to evaluate a website's adherence to Semantic Web standards and it proceeds to draw conclusions regarding how that integration affects their popularity in the modern competitive landscape of the Web.

Keywords: semantic web; media; art; culture; quantitative analysis; internet statistics; world wide web

Citation: Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Kouretsis, A.; Lamprogeorgos, A.; Varlamis, I. Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets. *Future Internet* **2022**, *14*, 36. https://doi.org/10.3390/fi14020036

Academic Editor: Rafael Valencia-Garcia

Received: 31 December 2021
Accepted: 17 January 2022
Published: 24 January 2022

1. Introduction

The Semantic Web, as a means to structure and disseminate data through machine-readability [1], is very important in the fields of art and culture and a cornerstone of digitized art and cultural heritage collections around the globe [2]. It is also a useful tool in the field of contemporary journalism, displaying enormous potential for information gathering, filtering and dissemination [3,4]. Since art and culture themselves are often the subject of journalistic content, the usage of Semantic Web technologies especially in media outlets that focus on these specific fields presents a very interesting landscape for research.

In the study presented in this article, the researchers proceeded to get a thorough glimpse at the landscape of Semantic Web information provided by art and culture-related websites with an added focus on the reportorial or journalistic aspects of this information. In order to do so, a variety of relevant websites were identified using a process that involved both automated and expert manual selection. These identified websites were then perused by an automated crawling algorithm and metrics about their integration of Semantic Web technologies were collected. The specific metrics selected were based on the researchers' expertise in the field of the Semantic Web and its application in art and cultural heritage and are presented in detail in the methodology section. Moreover, through this expertise an integration rating system was devised and is presented in detail. The information collected was further analyzed through means of not only traditional statistical analysis, but also a machine learning technique known as Gradient Boosting.

The ultimate goal of the study is to present an informed impression of the integration of various Semantic Web technologies in art and culture-related online media and assess both the perceived importance of these technologies and, to an extent, a quantitative measure of that importance in relation to a website's popularity.

2. Theoretical Background

More than two decades have passed since Tim Berners-Lee and his team envisioned the Semantic Web, a form of web content that would be readable and, thus, understandable and comprehensible by machines [1]. One of Semantic Web's most important properties is this ability to give more valuable information by automatically searching the meaning structure of web content [5]. In essence, it would provide structure to the anarchic structure of the Web in conjunction with the focus on human-centered computing as presented by Michael Dertouzos [6], an early proponent of the Semantic Web. The Semantic Web, alongside other web paradigms such as the Social Web which involves the evolution of social media, the 3D Web which encompasses virtual reality experiences in the Web and the Media Centric Web which focuses on the transmediative nature of the Web, is a key consisting element of Web 3.0 [7]. Furthermore, new technologies such as Artificial Intelligence and the blockchain also enhance and improve aspects of the Web, aiming to achieve different goals such as decentralization, connectivity and ubiquity. Web 3.0 and especially the Semantic Web, which is the focus of this study, seem to find their way in multiple thematic fields in the Web, such as news outlets or art and culture-related websites. The rise of increasingly technologically advanced forms of journalism dispels any questions about the technical optimization of contemporary journalism [3].

Despite the fact that now there is more pluralism in terms of websites than ever and more opinions can be heard, the sheer reality is that the majority of this information is left unstructured [3]. Semantic Web solutions could be valuable for journalistic research. The Web offered journalists the plurality that was missing but as a result this led to a more time-consuming process where journalists need to navigate through all the available data and sources and filter the information they are accessing manually [4]. Semantic Web technologies could automatically read, comprehend and include or exclude the useful information and even improve it by reproducing it with added enhancements such as accessibility features [8], or even advanced user personalization features [9].

Heravi and McGinnis [4] note that a combination of technologies will be necessary to provide a Social Semantic Journalism Framework. These technologies would undoubtedly collaborate with each other and serve as inputs and outputs for one another, establishing a procedure capable of addressing the issue that social media poses to journalists and editors as they attempt to determine what is noteworthy in user-generated content.

Another field that could benefit from Semantic Web solutions is that of art and cultural heritage in general. Cultural heritage can be defined as a kind of inheritance to be preserved and passed down to future generations. It is also linked to group identity and is both a symbol of and an essential ingredient in the building of that group's identity [10]. Cultural heritage is vital to understanding earlier generations and the origins of humanity. The Web has enabled local, national and global publishing, explanation and debate.

More and more museums, galleries and art-related institutions are transferring part or all of their collections into the digital space. The quality of a museum's Web presence on the Web can lead, not only to increased website visits, but also to increased physical visitors [11]. However, cultural heritage resources are vast and diverse. They comprise data or information that is highly structured, very unstructured or semi-structured and derived from both authorized and unauthorized sources, and also include multimedia data such as text, images, audio and video data [2]. In order to accommodate the users' needs, many usability-related features need to be implemented [12], but a usability-oriented approach is not the only approach that can help scientists, companies, and schools better understand cultural data. In the world of the Web, the data architecture of digital museum databases is quite diverse and calls for advanced mapping and vocabulary integration. This does

not come as a surprise as libraries, museums and galleries have always had extended and heterogeneous databases in the physical world as well. Several efforts have been conducted in recent years to digitize cultural heritage assets using Semantic Technologies such as RDF and OWL. There are numerous digital collections and applications available today that provide immediate access to cultural heritage content [13]. Additionally, cultural heritage is moving into new media of linear and non-linear storytelling, using audiovisual hypermedia assets and efforts are being made to provide enhanced semantic interaction, in order to transition such content into the Semantic Web [14].

Since the Web is a space full of available information, it goes without saying that someone can find available many sources related to art and culture that do not belong to their organizations but to websites, blogs or platforms focusing on arts and culture. In fact, art or cultural journalism is a distinctive field of journalism. Janssen [15] in his study about coverage of the arts in Dutch newspapers between 1965–1990, divided it in three levels: The first level regards the general newspapers' general portrayal of art, for example, the amount of space devoted to the arts in comparison to other topics. The second level examines disparities in the amount of focus provided to various creative forms or genres by contrasting, for example, classical and rock music coverage. The third level deals with the coverage that artifacts belonging to a certain artistic genre or subfield receive, for instance, the critical response to freshly released films. There was also a classification between cultural artifacts. The first level concerns the standing of the arts in respect to other (cultural) domains; the second level concerns the hierarchical relationships between art forms or genres; and the third level concerns the ranking of works and producers within a particular artistic domain.

Towards realizing their vision for the Semantic Web, the Semantic Web initiative of the World Wide Web Consortium (W3C) has developed a set of standards and tools to support this. Their early work resulted in two significant proposals: the Resource Description Framework Model and Syntax Specification and the Resource Description Framework Schema Specification. The W3C consisted of two primary working groups, the RDF Core Working Group and the Web Ontology Working Group, both of which issued significant sets of recommendations [16]. Since its inception, the Semantic Web has been evolving a layered architecture. Although there have been many variations since, its various components are:

- Unicode and URIs: Unicode as the computer character representation standard, and URIs, as the standard for identifying and locating resources (such as Web pages), offer a foundation for representing characters used in the majority of the world's languages and identifying resources.
- XML: XML and its relevant standards, such as and Schemas and Namespaces, are widely used for data organization on the Web, but they do not transmit the meaning of the data.
- Resource Description Framework: RDF is a basic information (metadata) representation framework that utilizes URIs to identify Web-based resources and a graph model to describe resource relationships. RDF lays the foundation for publishing and linking data. There are a number of syntactic representations available, including a standard XML format.
- RDF Schema: a simple type modelling language for describing classes of resources and properties between them in the basic RDF model. It provides a simple reasoning framework for inferring types of resources.
- Ontologies: a richer language capable of expressing more complicated constraints on the types and attributes of resources.
- Logic and Proof: an (automated) reasoning system built on top of the ontology structure with the purpose of inferring new relationships. Therefore, a software agent can deduce whether a certain resource fits its needs by utilizing such a system (and vice versa).

- Trust: This component is the final layer of the stack, and it refers to the issues of trust and trustworthiness of the information [16]. There are two main approaches regarding trust, the one is based on policy and the second on reputation [17]. Nowadays technology, trust and proof are regarded as the most emerging research areas of the Semantic Web [17].

Semantic Web technologies are regarded as an approach to manage knowledge by utilizing ontologies and semantic web standards, allow individuals to establish data repositories on the Web, create vocabularies, and write rules for data processing. Linked data are assisted by technologies such as RDF, SPARQL, OWL and SKOS [18]. Additionally, a very classic but perhaps outdated framework is the Ranking Semantic Search framework (RSS). This framework enables ranking of the search results on the Semantic Web and, through the use of novel ranking strategies, avoids returning disordered search results [19].

Semantic Web technologies can be applied to a website so it will be more easily readable and accessible by search engines for better Search Engine Optimization (SEO). For instance, website owners or content managers can enhance their text descriptions with semantic annotations and check if this leads to a more satisfying user experience. Towards this end, Necula et al. [20] investigated e-commerce websites and whether there is a correlation between the enhancement of product text descriptions with semantic annotations and the perceived consumers' satisfaction. Their study concluded that the inclusion of semantic web elements in the products descriptions is important for a more pleasant customer experience. In fact, one of the most interesting findings was that the consumer regards knowledge graphs as having high significance in an e-commerce website [20].

A way to add information that is machine-understandable to Web pages that is processed by the major search engines to improve search performance is schema.org [21]. Schema.org's wide adoption is related to its promotion by major search engines as a standard for marking up structured data in HTML web pages [22]. This adoption addresses a fundamental issue for the Web, by making it easy to annotate data within websites, at least for the most common types of Web content [23].

Although using Semantic Web technologies will lead to a more pleasant and usable user experience, it is not certain that this automatically means an improvement in terms of popularity. It is a fact that SEO techniques are used in order to improve a website searchability and consequently popularity, but it is not certain that utilizing Semantic Web technologies will automatically result in increased popularity. This is what this research tries to shed light on.

3. Methodology

3.1. Relevant Website Discovery

In order to gain as much information as possible concerning the level of Semantic Web integration in art and culture-related online media, collecting a big sample of appropriate websites was an essential requirement. Identifying such websites was a complex process involving both automated procedures and human expert input, in order to achieve the best results. The study's website discovery process did not attempt to genuinely discover any and all existing appropriate websites, but instead focused on collecting a sufficiently large sample.

3.1.1. Automated Sampling

The first step in this process was to acquire an up-to-date list of websites belonging to the generic Top Level Domains (gTLDs) that any person or entity is permitted to register, which are the .com, .net and .org gTLDs. The rest of original gTLDs (.int, .edu, .gov and .mil) were excluded since the study's main focus was on private activity, both commercial and non-profit. Such a list was acquired through Common Crawl a "non-profit organization dedicated to providing a copy of the internet to internet researchers, companies and individuals" [24]. The acquired list concerned websites that were indexed in October 2021 thus making it appropriately relevant. In order to pinpoint websites that offered what the

study required, a series of keywords were used. These keywords were "art", "media", "entertain" as short for entertainment and "cult" as short for culture. This starting point of relevant website discovery procured 449,063 Second Level Domains (SLDs), as seen in Table 1.

Table 1. Initially collected SLD quantities.

gTLD	Art	Media	Entertain	Cult
.com	326,153	49,964	7023	10,578
.net	17,577	3195	366	663
.org	28,413	2574	125	2432
Totals	372,143	55,733	7514	13,673

In order to accomplish the above, an automated process was created. This process, which was developed in PHP, used Common Crawl's API to receive a list of domains page by page. Then it proceeded to filter out all subdomains and sub-folders and to check each domain name for the specified keywords. For domains that were available through both HTTP and the more secure HTTPS the non-secure version was filtered out. For domains available both with the www subdomain and without, the ones with www were excluded. If no secure version was available, the non-secure one was kept. The same principle was applied with regards to the www subdomain. This procedure's flowchart can be seen in Figure 1.

Since the websites would be required to be evaluated on their content in order to establish that they are indeed relevant to art or culture and include what may be considered reportorial content, any websites that did not support the English language were excluded. Dual language websites were accepted as long as English was the primary language. This decision was largely influenced by the fact that the English language is in a position of dominance in the Web compared to other languages [25,26]. This is directly related to the "digital divide"—the inequality present in the information society which stems from the difficulty of Internet access for a significant portion of humanity [26,27]. The language of a website was identified based on Common Crawl Index's language parameter. This reduced the number of potentially useful websites to 252,105 sites. Consequently, a number of these websites were filtered out based on the presence in their sTLD of irrelevant words that share part of them with the aforementioned keywords (i.e., earth, heart, quarter, smart, chart, part, etc.)

The next step in narrowing down the number of websites that had to be evaluated was identifying their popularity. Even though the World Wide Web is full of interesting art blogs, cultural publications, artist collectives and more, it is expected that the most popular websites are the ones that have the most impact and the ones worth focusing on. In order to assess each site's popularity, Alexa's Web Information Service was used. Alexa Internet is an Amazon.com company providing insight on site popularity for more than 25 years [28]. The Web Information Service provides information about web sites through the use of a Web services API. Part of this information is the rank of almost 10 million websites based on traffic in the last 90 days. An automated process that used Alexa's Web Information Service through API requests was implemented in order to gather information for all the websites in our database. After eliminating all low-traffic websites, a total of 16,616 websites remained.

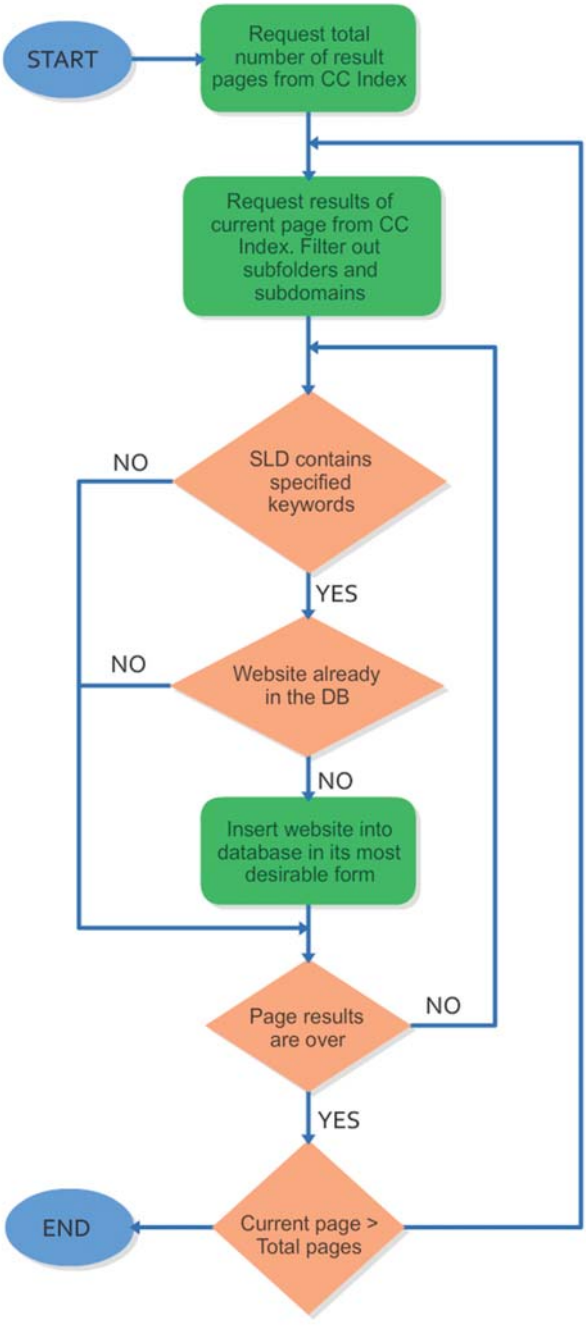


Figure 1. Flow chart of the website discovery crawler.

3.1.2. Expert Screening

In order to identify which of the remaining websites actually corresponded with the researcher's intends the rest of the screening was accomplished through manually visiting and browsing through the websites. This process was conducted by the members of the research team themselves. The Team consists of multiple experts with years of accumulated experience in studying the Web presence of Audiovisual Arts and Cultural Heritage.

In an effort to accommodate this intense and time-consuming process, a small-scale Web application was designed and implemented. Its purpose was to present the researchers with a number of potential websites alongside some information about them like their gTLD and the keyword that their SLDs contained. The members of the research team would then visit the website and after their audit they could choose to evaluate it by selecting the appropriate of three color-coded options:

- Red indicated that a website that had no relevance to the field of art and culture.
- Yellow indicated that a website was art or culture related but had limited reportorial content.
- Green indicated that a website was not only art or culture-related but also contained a fair amount of reportorial content.

The two main criteria for this evaluation were each website's relevance to the fields of art or cultural heritage and whether the website's content was even partially of a journalistic or reportorial nature. Websites that contained information about works of art, artists and their past and current projects, local or international culture and cultural heritage, cultural artifacts or places, historical or academic analysis of artworks, and so on, were considered by the researchers relevant to the fields of art or culture. Such websites included, but were not limited to, websites of museums, galleries, collections, art schools and colleges, artist portfolios, organizations or societies promoting art and culture, news outlets covering relevant matters, artist agencies, art vendors and more. Any non-relevant websites were marked as "Red" and excluded from the study. The researchers also searched for reportorial content inside each website such as articles, blog entries, opinion pieces, artwork analyses, news regarding events or exhibitions, artwork reviews, artist interviews, historical retrospects, current artistic event discussion and more. Websites that exhibited a fair amount of such reportorial content were evaluated as "Green" while those that were relevant to art but exhibited extremely limited or no reportorial content were evaluated as "Yellow".

Figure 2 presents a screenshot of the Web application during its use. The interface also presents the total number of evaluations required, as well as the current number of evaluations completed by this researcher. Additionally, it presents a small preview of how the evaluation process is shaping up by indicating how many websites have been so far evaluated in each category.

Out of a total of 16,616 websites, 12,874 were evaluated as Red, 2653 were evaluated as Yellow and 1089 were evaluated as Green. In addition, the researchers were encouraged to suggest additional websites that they knew fit the criteria and were not discovered by the automated process. This led to an additional 35 websites that were added to the pool and evaluated as Green, bringing the total number of evaluated websites to 16,651 and the total value of Green websites to 1124.

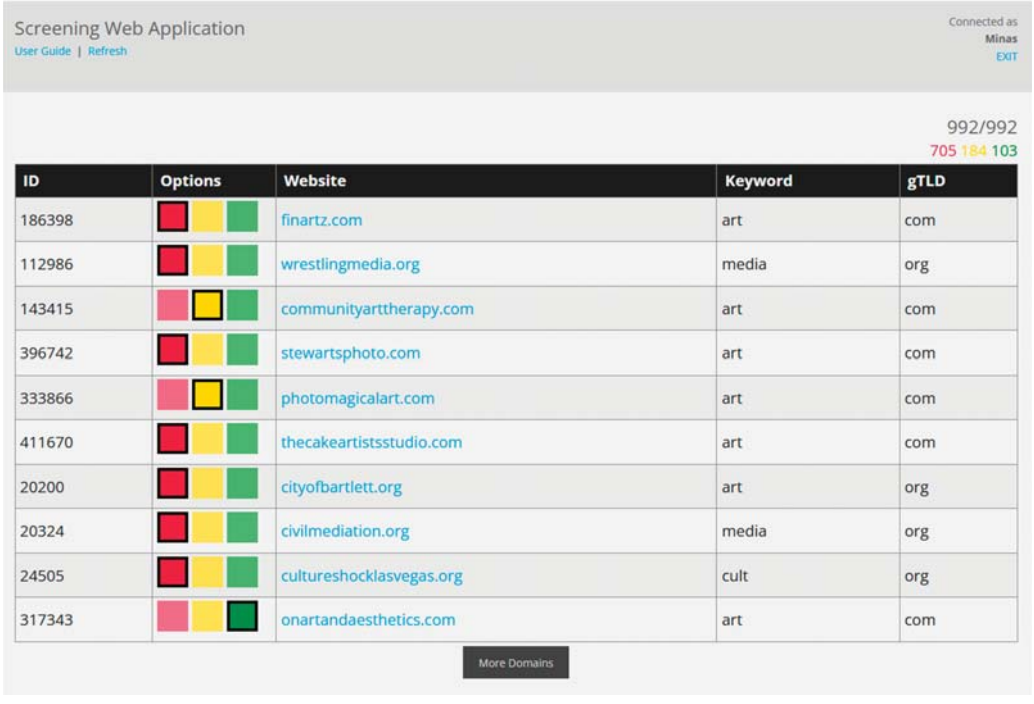


Figure 2. Interface of the Web application created to facilitate the manual screening process.

3.2. Collecting Information

The next step of the study involved investigating the relevant websites in order to collect information regarding which Semantic Web technologies were integrated in each website and, where possible, to what extent. As a means of accomplishing this, an automated procedure was developed and implemented in PHP making use of the cURL Library, a library created by Daniel Stenbergand to allow for connectivity and communication with different servers through various protocols [29], and the DOM PHP extension. All websites evaluated as Yellow and Green were deemed important for this step, since a website is also an information outlet in and of itself. As a result, 3777 websites were investigated.

This procedure connected to the websites' homepage and identified all internal links presented there. It then proceeded to "crawl" through these links and attempted to detect the use of various specific methods that had as a goal to assist with each website's machine-readability. For every website a maximum of 80 pages, including the homepage, were crawled, in an effort to avoid spending an overly extended time in a single website. This number of pages was deemed by the researchers capable of providing a comprehensive impression of the extent of integration of Semantic Web technologies. After identifying these technologies, the crawler also attempted to extract metrics on their usage in manners that will be further elaborated upon below. Out of the 3777 websites, 3632 were successfully crawled. The unsuccessful attempts included websites that denied automated indexing through a robots.txt file or where the crawler encountered various technical difficulties. Figure 3 presents the flow chart of the crawler.

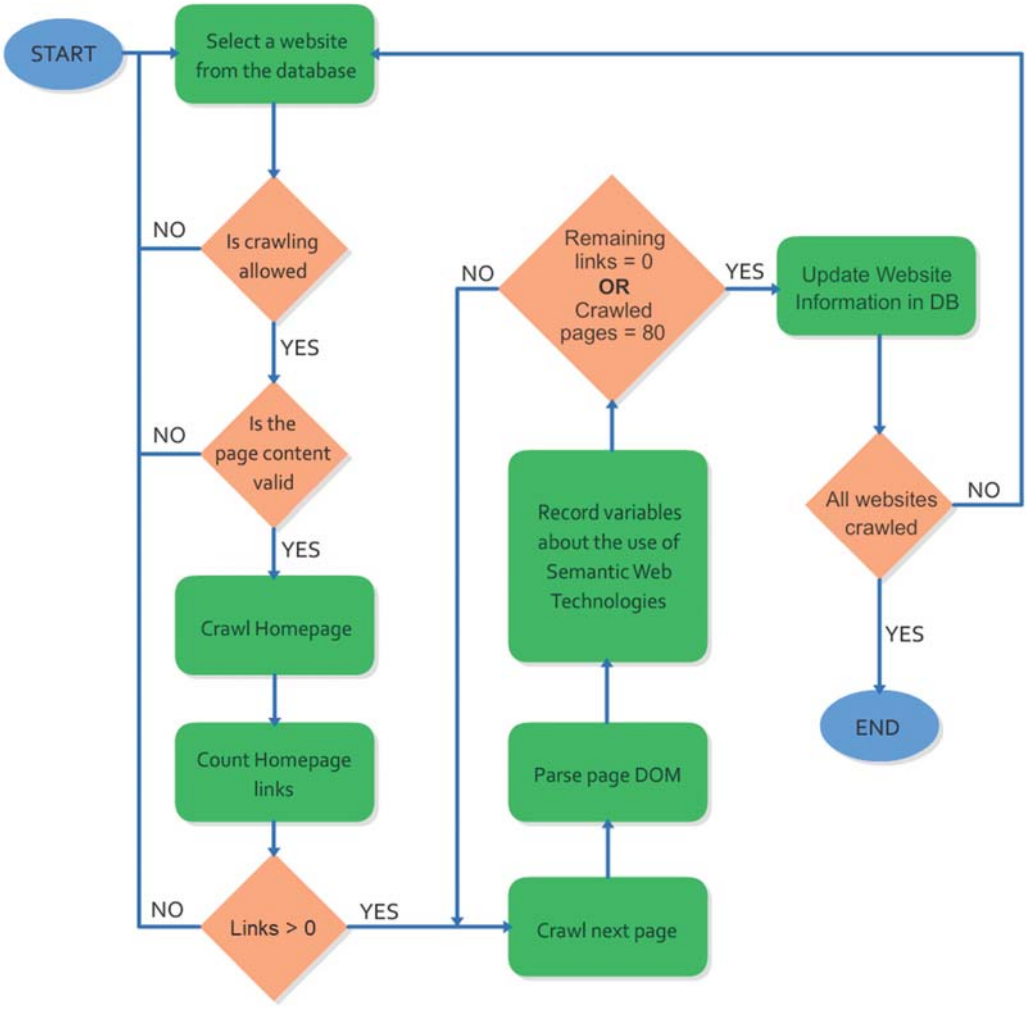


Figure 3. Flowchart of the Semantic Web technologies detection crawler.

The Semantic Web technologies investigated were:

- The use of RSS feeds;
- The use of HTML5 Semantic Elements;
- The use of the Open Graph protocol;
- The use of Twitter Cards markup;
- The use of schema.org schemas;
- The use of Microformats metadata format.

These different methods of creating a more machine-readable website are detailed below.

3.2.1. RSS Feeds

RSS (RDF Site Summary) is a format that allows the creation of Web feeds [30] that can be used to allow applications to access a website's information. It is one of the earliest attempts at Web syndication and through its popularity in the 2000s it stood in the forefront

of creating a machine-readable Web. The study's algorithm detected how many unique RSS feeds were available in each individual website (variable _rss_feeds).

3.2.2. HTML Semantic Elements

The use of new Semantic Elements was introduced in HTML5 in an effort to help define the content of various structural elements of the Document Object Model (DOM) not only to the developer but also to the browser [31]. These are specifically the elements <article>, <aside>, <details>, <figcaption>, <figure>, <footer>, <header>, <main>, <mark>, <nav>, <section>, <summary> and <time>. The study's crawling algorithm located such elements in a Web page's structure and counted how many pages of each website included these elements (variable _html). Additionally, it monitored how many different such elements were used for each individual website (variable _html_variety).

3.2.3. Open Graph

The Open Graph protocol contains metadata that help create a rich object regarding each Web page for the purpose of displaying it in a social graph [32]. The protocol follows a method compatible with W3C's RDFa (Resource Description Framework in attributes) recommendation. The most basic metadata element of the protocol is the og:title element which contains a title for the Web page as it would appear on the graph. The study's algorithm detected how many pages of each website included an og:title element (variable _og). Additionally, it monitored what percentage of these titles where unique to a single specific page and not a reused generic title (variable _og_variety).

3.2.4. Twitter Cards

Twitter Cards use a markup system to create a rich object specifically for the Twitter social media platform [33]. Similarly to the Open Graph system, it complies with RDFa syntax. The Summary card is a twitter card which creates a short summary of the specific Web page. The title of that summary can be located in the twitter:title metadata element. The study's algorithm detected how many pages of each website included a twitter:title element (variable _twitter). Additionally, it monitored what percentage of these titles where unique to a single specific page and not a reused generic title (variable _twitter_variety).

3.2.5. Schema.org Schemas

Schema.org is a community-driven collection of structured data schemas [22] for use on the Internet, founded by Google, Microsoft, Yahoo and Yandex. Its purpose is to make it easier on website developers to integrate machine-readable data in their websites. The data can be conveyed using different encodings such as RDFa, Microdata or JSON-LD. The study's algorithm detected how many pages of each website included a schema.org element in any of these three different methods of encoding (variable _schemaorg).

3.2.6. Microformats

Microformats is a set of data formats that can be used to convey machine-readable data. The various data formats are explicitly declared through the use of HTML Classes [34]. Multiple such Microformats are available in order to semantically mark a variety of information. The study's algorithm detected how many subpages of each website included one of the following classes indicated usage of a microformat: "adr", "geo", "hAtom", "haudio", "vEvent", "vcard", "hlisting", "hmedia", "hnews", "hproduct", "hrecipe", "hResume", "hreview", "hslice", "xfolkentry" and "xoxo" (variable _microformats). Additionally, it monitored how many different such classes were used for each individual website (variable _microformats_variety).

In addition to the above technologies, the crawling algorithm developed in this study kept a record on how many pages of each website were crawled (variable _pages_crawled) as well as any other json + app formats that might appear in a page that might be worth

investigating (variable _other). A comprehensive table of all variables recorded by the study's algorithm is presented in Table 2.

Table 2. List of SWT related variables recorded by the crawler algorithm.

Variable	Short Description
_pages_crawled	Number of pages crawled
_rss_feeds	Number of unique RSS feed links
_html	Number of pages with HTML5 Semantic Elements
_html_variety	% of HTML5 Semantic Elements used
_og	Number of pages with Open Graph Metadata Elements
_og_variety	% of og:title values that are unique
_twitter	Number of pages with Twitter Summary Card Metadata Elements
_twitter_variety	% of twitter:title values that are unique
_schema_org	Number of pages with schema.org structured data
_microformats	Number of pages with Microformats data formats
_microformats_variety	% of Microformats used
_other	Number of pages with other json data

3.3. Evaluating Semantic Web Technologies Integration

The multitude of measured quantitative variables that were collected during the website crawling process are all indicators of a website's adherence to Semantic Web standards. They can be used to get a glimpse of how committed each website is to making its information machine-readable. As part of the effort of documenting this commitment the researchers have created a 5-star rating system that can translate the measurements in an easy-to-read comprehensive value dubbed "Semantic Web Technologies Integration Rating" or SWTI rating.

The rating system focuses on which elements the researchers consider most important through their expertise in the field of integration of Semantic Web technologies.

The usage of structured data is the first and most important aspect of such an integration. Schema.org is supported by multiple colossi of the Web such as Google and Microsoft and Microformats has a long history of effort in promoting machine-readability. Hence, one star is rewarded for attempting use of these technologies at least to some extent, with a second star being rewarded to websites that have a more extensive integration.

The creation of rich objects for social media may stem from a different motivation but nonetheless it is a major contributing factor in the machine-readability of modern websites. As such, one star is awarded for the implementation of at least one such method, either Open Graph or Twitter Cards. An additional half star is awarded if the implementation focuses in providing unique information for each different page of a website, as dictated by usage guidelines.

The use of HTML5 semantic elements promotes a content-based structure of a website's DOM at least to some extent and so it is rewarded with half a star. Additionally, when the website uses multiple different such elements it becomes an indicator of a quality implementation and as such it is rewarded with another half star.

Finally, providing an RSS feed has been a popular practice for more than 20 years and it is a good first step in assisting with machine-readability. Since RSS popularity is declining, its use awards half a star.

The scoring system is presented in detail in Table 3.

Table 3. SWTI rating system.

Condition	Stars Awarded
Website using at least some schema.org or Microformats structured data.	1
Website using Schema.org or Microformats structured data in at least 50% of its crawled pages	1
Website using Open Graph or Twitter Cards to provide at least one rich object for social media	1
Website using unique Open Graph or Twitter Card titles for over 50% of its rich objects	0.5
Website using at least one HTML5 Semantic Element	0.5
Website using at least 50% of the different HTML5 Semantic Elements	0.5
Website providing at least one RSS feed	0.5
Total	5

4. Results

4.1. Statistics and Analysis

Table 4 depicts a sample of the first ten entries of our data formation. The first column shows the websites, the second column presents the ranking of a website based on its popularity according to Alexa Internet (_alexa_rank). Columns 3 to 7 contain scores for each of the four major Semantic Web technologies derived by dividing the pages that used each technology as shown in Section 3.2 (variables _rss_feeds, _html, _og, _twitter, _schema_org) by the total number of pages crawled (variable _pages_crawled) thus creating the new variables (_rss_score, _html_score, _og_score, _twitter_score, _schema_score). Columns from 8 to 10 contain the variables _html_variety, _og_variety and _twitter_variety. The last column contains the rating for each site based on the SWTI rating system detailed in Section 3.3 (variable _swti). The variables _microformats, _microformats_variety and _other, which identified usage of microformats or other json data in each web page were omitted from further statistical analysis because the percentage of websites with findings in these metrics was below 1%.

Table 4. Data formation sample.

Domain	Alexa Rank	RSS Score	Html Score	OG Score	Twitter Score	Schema Score	Html var%	OG var%	Twitter var%	SWTI Rating
03mediainc.com	228,588	60.00	95.00	95.00	0.00	95.00	77	97	0	5.00
10xplusmedia.com	269,411	38.46	96.15	96.15	0.00	96.15	31	96	0	4.50
13artists.com	4,084,650	27.27	90.91	90.91	0.00	81.82	23	100	0	4.50
1531entertainment.com	1,278,072	62.50	100.00	100.00	100.00	0.00	77	100	70	3.00
1913mediagroup.com	4,799,977	60.00	100.00	0.00	0.00	0.00	31	0	0	1.00
1artworks.com	4,141,675	65.22	91.30	91.30	86.96	0.00	31	100	5	2.50
1atbatmedia.com	2,355,870	47.06	100.00	100.00	0.00	100.00	54	94	0	5.00
1media-en.com	2,216,622	101.25	71.25	98.75	0.00	98.75	46	97	0	4.50
03mediainc.com	228,588	30.00	17.50	85.00	0.00	85.00	31	100	0	4.50
10xplusmedia.com	269,411	0.00	100.00	96.25	0.00	100.00	23	99	0	4.00

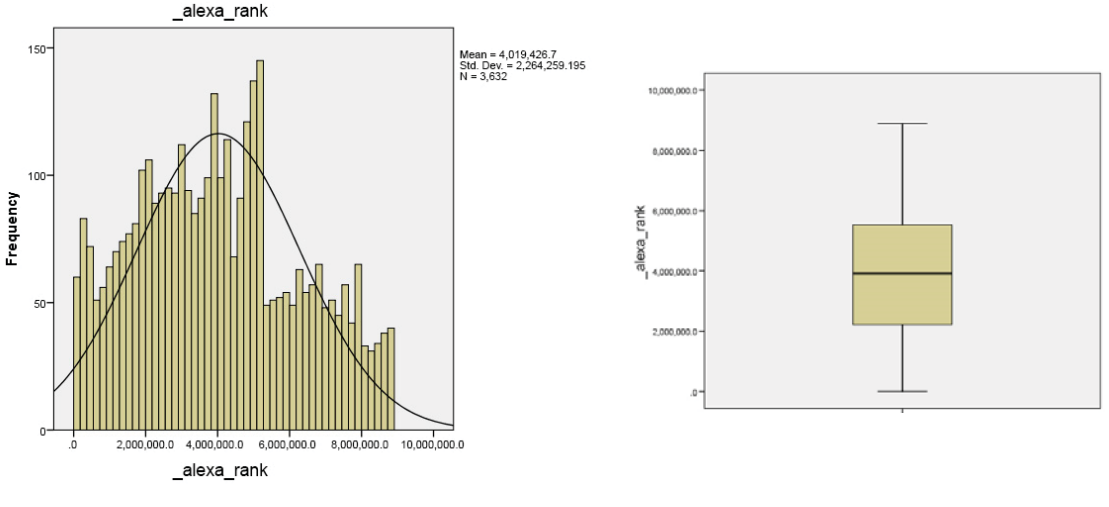
Table 5 depicts the descriptive statistics for each variable and Table 6 depicts the frequency related statistics. In Figure 4 the histogram and boxplot of the _alexa_rank variable are presented, depicting the distribution and dispersion of the variable. The histogram and boxplot, the distribution and dispersion of the other sample values are plotted for each of the variables and presented in Appendix A.

Table 5. Descriptive statistics.

	N	Minimum	Maximum	Mean	Std. Deviation
_alexa_rank	3632	3736.0	8,887,606.0	4,019,426.703	2,264,259.1955
_rss_score	3632	0.0	300.0	20.325	33.6279
_html_score	3632	0.0	100.0	77.128	35.1675
_og_score	3632	0.0	100.0	65.814	41.7498
_twitter_score	3632	0.0	100.0	31.371	43.5979
_schema_score	3632	0.0	100.0	42.751	44.2935
_swti	3632	0.0	5.0	2.968	1.5851
_html_variety	3632	0.0	92.0	36.697	21.3882
_og_variety	3632	0.0	100.0	70.367	42.8166
_twitter_variety	3632	0.0	100.0	33.598	44.1914
Valid N (listwise)	3632				

Table 6. Frequencies.

		_alexa _rank	_rss _score	_html _score	_og _score	_twitter _sctore	_schema _score	_html _variety	_og _variety	_twitter _variety	_swti
N	**Vald**	3632	3632	3632	3632	3632	3632	3632	3632	3632	3632
	Mis	0	0	0	0	0	0	0	0	0	0
Mean		4,019,426.70	20.325	77.128	65.814	31.371	42.751	36.697	70.367	33.598	2.968
Std. Err. of Mean		37,571.0403	0.5580	0.5835	0.6928	0.7234	0.7350	0.3549	0.7105	0.7333	0.0263
Std. Deviation		2,264,259.19	33.6279	35.1675	41.7498	43.5979	44.2935	21.388	42.8166	44.1914	1.5851
Variance		5.127×10^{12}	1130.837	1236.75	1743.047	1900.77	1961.915	457.45	1833.257	1952.879	2.513
Skewness		0.225	2.702	−1.476	−0.777	0.772	0.226	−0.177	−0.949	0.634	−0.315
Kurtosis		−0.808	11.401	0.534	−1.206	−1.309	−1.810	−0.703	−1.003	−1.501	−1.190
Range		8,883,870.0	300.0	100.0	100.0	100.0	100.0	92.0	100.0	100.0	5.0
Minimum		3736.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Maximum		8,887,606.0	300.0	100.0	100.0	100.0	100.0	92.0	100.0	100.0	5.0
Percent	**25**	2,210,762.75	0.000	75.000	5.212	0.000	0.000	23.000	7.000	0.000	1.500
	50	3,915,896.50	3.101	95.000	90.909	0.000	20.000	38.000	100.000	0.000	3.000
	75	5,530,911.25	28.571	100.000	100.000	90.183	94.118	54.000	100.000	89.000	4.500



(**a**) (**b**)

Figure 4. The histogram (**a**) and boxplot (**b**) of the _alexa_rank variable showing a relatively normal distribution.

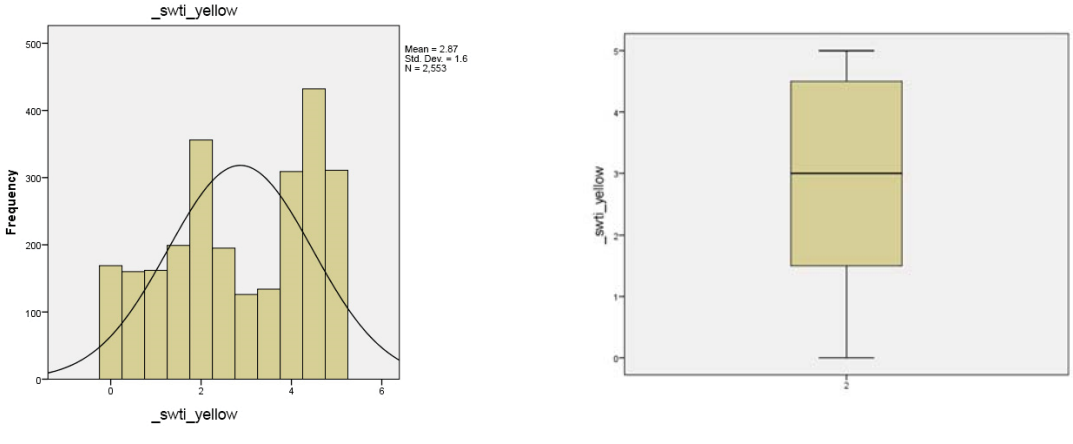
Table 7 depicts the descriptive statistics of the _swti variable for websites that were evaluated as Yellow or Green during the expert screening process described in Section 3.1.2 (variables _swti_yellow, swti_green). Frequency-related statistics for these variables are presented in Table 8 and their histograms and boxplots in Figures 5 and 6.

Table 7. Descriptive statistics for the SWTI of Yellow and Green websites.

	N	Minimum	Maximum	Mean	Std. Deviation
_swti_yellow	2553	0	5	2.87	1.600
_swti_green	1079	0	5	3.20	1.524
Valid N (listwise)	1079				

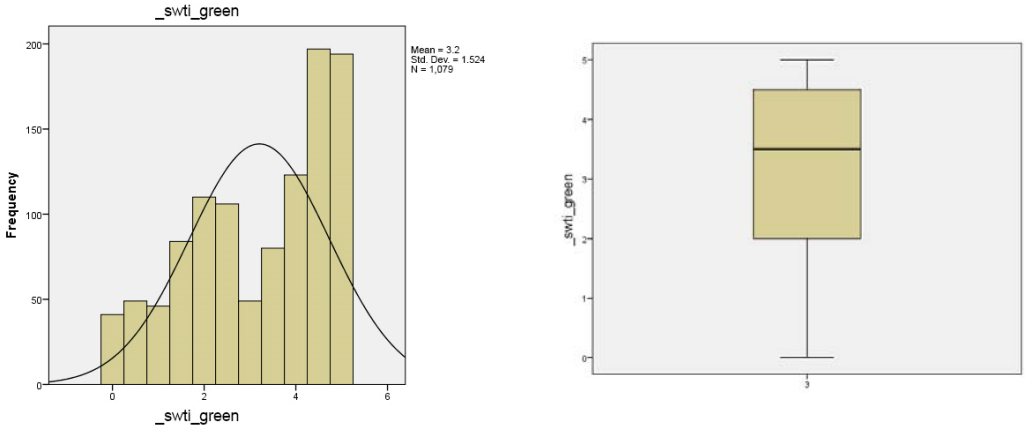
Table 8. Frequencies for the SWTI of Yellow and Green websites.

		_swti_yellow	_swti_green
N	**Valid**	2553	1079
	Missing	1079	2553
Std. Error of Mean		0.032	0.046
Std. Deviation		1.600	1.524
Variance		2.561	2.321
Skewness		−0.244	−0.483
Std. Error of Skewness		0.048	0.074
Kurtosis		−1.245	−0.997
Std. Error of Kurtosis		0.097	0.149
Range		5	5
Minimum		0	0
Maximum		5	5
Percentiles	**25**	1.50	2.00
	50	3.00	3.50
	75	4.50	4.50



(**a**) (**b**)

Figure 5. The histogram (**a**) and boxplot (**b**) of the _swti_yellow variable showing a non-normal distribution.



(**a**) (**b**)

Figure 6. The histogram (**a**) and boxplot (**b**) of the _swti_green variable showing a non-normal distribution.

In order to analyze the interrelation between the independent variables _html_score, og_score, _twitter_score and _schema_score, a Pearson's r criterion was applied [35]. The results are shown in Table 9 where the correlations between the variables are depicted. All the correlations have significance level at 0.000 confirming the statistically significant correlation.

Table 9. Pearson correlations between html_score, og_score, twitter_score and schema_score.

		html_score	og_score	twitter_score	schema_score
html_score	**Pearson correlation**	1	0.393 **	0.316 **	0.349 **
	sig.		0.000	0.000	0.000
	N	3632	3632	3632	3632
og_score	**Pearson correlation**	0.393 **	1	0.482 **	0.422 **
	sig.	0.000		0.000	0.000
	N	3632	3632	3632	3632
twitter_score	**Pearson correlation**	0.316 **	0.482 **	1	0.099 **
	sig.	0.000	0.000		0.000
	N	3632	3632	3632	3632
schema_score	**Pearson correlation**	0.349 **	0.422 **	0.099 **	1
	sig.	0.000	0.000	0.000	
	N	3632	3632	3632	3632

** Correlation is significant at the 0.01 level.

The Pearson's r coefficients range from 0.316 (the weaker positive correlation between html_score and twitter score) to 0.482 (the stronger positive correlation between og_score and twitter score).

In an effort to examine the interrelationship between the collected Semantic Web metrics and a websites popularity the various websites were ranked according to their measured SWTI rating (variable _swti_rank) and the Spearman's rank correlation coefficient was calculated. The results are presented in Table 10.

Results of the Spearman correlation indicated that there is a significant very small positive relationship between _swti_rank and _alexa_rank and Y, ($r(3630) = 0.0683$, $p < 0.001$). This correlation despite being very small prompted the researchers to further investigate the interrelationship between SW integration and popularity using a gradient boosting analysis which included every metric collected by the crawling algorithm.

Table 10. Spearman correlation between _swti_rank and _alexa_rank.

Parameter	Value
Spearman correlation coefficient (r)	0.06835
***p*-value**	0.0000375
Covariance	161,169,282.6
Sample size (n)	3632
Statistic	4.1275

4.2. Gradient Boosting Analysis Using XGBoost

After samples have been collected, the XGBoost models are built using a grid search among the parameter space. XGBoost (eXtreme Gradient Boosting) is a fast implementation of gradient boosting [36]. It is a scalable end-to-end tree boosting system that has been widely used and achieves state-of-the-art classification and regression performance [37]. It can improve in the reduction of overfitting, the parallelization of tree construction, and the acceleration of execution. It is an ensemble of regression trees known as CART [38]. The prediction score is calculated by adding all of the trees together, as indicated in the following equation,

$$\hat{Y} = \sum_{m=1}^{M} f_m(X) \tag{1}$$

where M is the number of trees and f_m is the mindependent CART tree. In contrast to Friedman's [39] original gradient boosting architecture, XGBoost adds a regularized objective to the loss function. The regularized objective for the mth iteration optimization is provided by

$$L^m = \sum_{i=1}^{n} l(y_i, \hat{y_l}^m) + \sum_{j=1}^{m} \Omega(f_j) \tag{2}$$

where n denotes the number of samples, l denotes the differentiable loss function that quantifies the difference between the predicted $\hat{y_i}^m$ and the target y_i and Ω denotes the regularization term

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{k=1}^{T} {w_k}^2 \tag{3}$$

where T is the number of nodes and w denotes each node's weight. The regularization degree is controlled by two constants, γ and λ. Furthermore, taking into account that for the mth iteration the following relation holds,

$$\hat{y_l}^m = \hat{y_l}^{m-1} + f_m(x_i) \tag{4}$$

we can recast Equation (2) as,

$$L^m = \sum_{i=1}^{n} \left[l\left(y_i, {y_i}^{m-1}\right) + g_i f_m(x_i) + \frac{1}{2} h_i {f_m}^2(x_i) \right] + \Omega(f_m) \tag{5}$$

where we introduced the operators, $g_i = \partial_{\hat{y}^{m-1}} l\left(y_i, \hat{y}^{m-1}\right)$ and $h_i = \partial^2{}_{\hat{y}^{m-1}} l\left(y_i, \hat{y}^{m-1}\right)$, which are the loss function's first and second-order derivatives, respectively.

XGBoost makes the gradient converge quicker and more accurately than existing gradient boosting frameworks by using the second-order Taylor expansion for the loss function [36]. It also unifies the generation of the loss function's derivative. Furthermore, adding the regularization term XGBoost to the target function balances the target function's decrease, reduces the model's complexity, and successfully resolves overfitting [36].

Furthermore, XGBoost can use the weight to determine the importance of a feature. The number of times a feature is utilized to partition the data across all trees is the weight in XGBoost [36], and is given by the equation

$$IMP^F = \sum_{m=1}^{M} \sum_{l=1}^{L-1} I\left(F^l{}_m,\ F\right) I\left(F^l{}_m,\ F\right) \tag{6}$$

with the boundary conditions, $IMP^F = 1$, if $F^l{}_m == F$, else $IMP^F = 0$. M is the number of trees or iterations, L denotes the number of nodes in the mth tree, $L-1$ denotes the tree's non-leaf nodes, $F^l{}_m$ stands for the corresponding feature to node l, and $I()$ denotes the indicator function.

The Alexa ranking for the websites under investigation is used as the outcome of the fitted model. The features collected by the crawling mechanism are used as the predictor variables. Since the main point of the analysis is to identify the most important features related to semantic web technologies with respect to the ranking of a website, we perform a grid search for the parameter space of XGBoost. The Alexa ranking is used to extract four classes using the quartiles with respect to the ranking. This transforms the regression analysis to a multiclass classification problem with four classes available. The first class is for the top 25% of the websites in ranking, and the other three classes are for the intervals [0%, 25%), [25%, 50%) and [50%, 75%] of the remaining websites.

The measure logLoss, or logarithmic loss, penalizes a model's inaccurate classifications. This is particularly useful for multiclass classification, in which the approach assigns a probability to each of the classes for all observations (see, e.g., [40]). As we are not expecting a binary response, the logLoss function was chosen over traditional accuracy measurements. The logLoss function is given by

$$logLoss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} ln(p_{ij}) \tag{7}$$

where, M is the number of classes, N the number of observations, $y_{ij} = \{0,\ 1\}$ indicates if observation i belongs to class j, an p_{ij} the respective probability.

The number of pages crawled is used to scale the related features extracted. These are all page count features for a respective semantic web technology, and the feature extracted for the rss feeds. In addition, this transformation "scales-out" the number of pages crawled to isolate the effect, and the importance of the semantic web features measured to ranking. In particular, the following variables are transformed by dividing with the number of pages crawled ("_pages_crawled"), "_html", "_og", "_twitter", "_rss_feeds", "_schema_org", "_other", "_microformats".

The parameters of machine learning models have a significant impact on model performance. As a result, in order to create an appropriate XGBoost model, the XGBoost parameters must be tuned. XGBoost has seven key parameters: boosting number (or eta), max depth, min child weight, sub sample, colsample bytree, gamma, and lambda. The number of boosting or iterations is referred to as the boosting number. The greatest depth to which a tree can grow is represented by max depth. A larger max depth indicates a higher degree of fitting, but it also indicates a higher risk of overfitting. The minimum sum of instance weight required in a child is called min child weight. The algorithm will be more conservative if min child weight is set to a large value. The subsample ratio of the training instances is referred to as subsample. Overfitting can be avoided if this option is set correctly. When constructing each tree, colsample bytree refers to the subsample ratio of features. The minimum loss reduction necessary to make a further partition on a tree leaf node is referred to as gamma. The higher the gamma, the more conservative the algorithm is. Lambda represents the L2 regularization term on weights. Additionally, increasing this value causes the model to become more conservative. We perform a grid search using the facilities of the Caret R-package [41]. We search the parameter space with the "grid" method, using 10-fold cross validation for a tuneLength of 30 that specifies the total number of unique combinations using the trainControl and train functions of the Caret package (Kuhn 2008). The optimal values identified are

{eta = 0.3, gamma = 0, min child weight = 5, max depth = 6, subsample = 0.5, colsample_bytree = 0.5, lambda = 0.5}.

The overall statistics are presented in Table 11 and the statistics by class in Table 12.

Table 11. Overall statistics.

Overall Statistics	
Accuracy	0.304
95% CI	(0.2742, 0.335)
***p*-Value**	0.0009703
Kappa	0.0727

Table 12. Statistics by class.

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.3259	0.23504	0.4081	0.25110
Specificity	0.7968	0.79228	0.6905	0.79295
Pos Pred Value	0.3443	0.28205	0.3003	0.28788
Neg Pred Value	0.7830	0.74895	0.7818	0.76056
Precision	0.3443	0.28205	0.3003	0.28788
Recall	0.3259	0.23504	0.4081	0.25110
F1	0.3349	0.25641	0.3460	0.26824
Prevalence	0.2467	0.25771	0.2456	0.25000
Detection Rate	0.0804	0.06057	0.1002	0.06278
Detection Prevalence	0.2335	0.21476	0.3337	0.21806
Balanced Accuracy	0.5613	0.51366	0.5493	0.52203

Figure 7 presents the sorted accuracy for each model fit and Figure 8 displays the various variables and their importance.

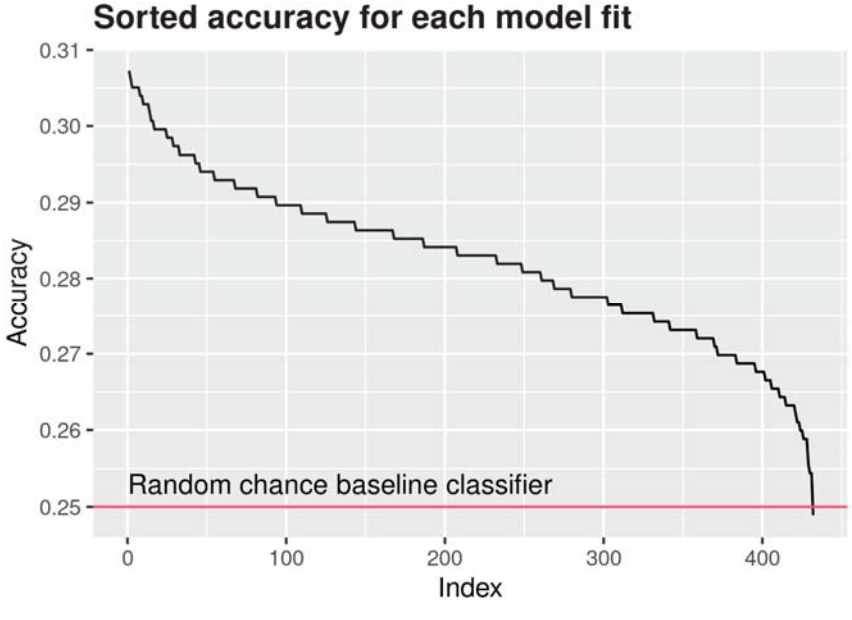


Figure 7. Sorted accuracy for each model fit.

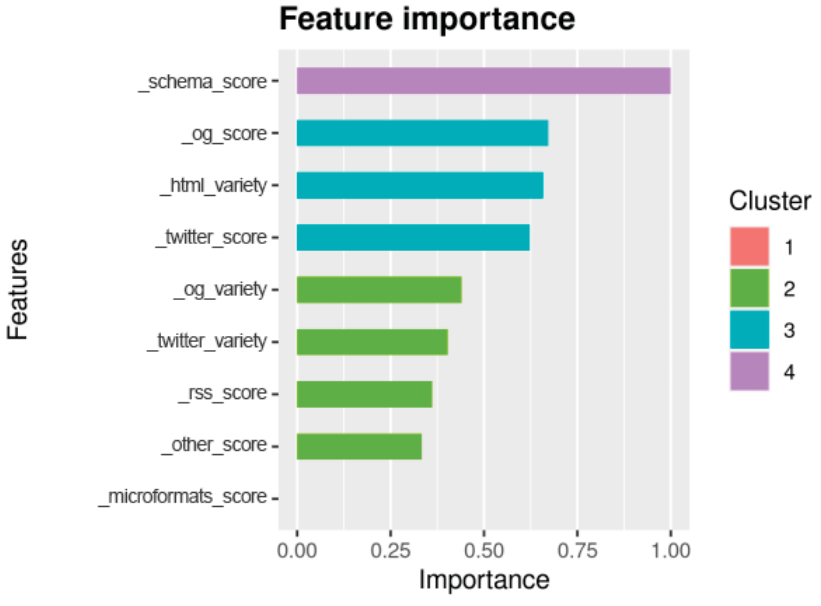


Figure 8. Feature importance.

5. Discussion

The metrics presented in Section 4.1 provide an interesting overview of the landscape of Semantic Web technologies integration in media websites with content relevant to art and cultural heritage. There are some common patterns that seem to emerge while at the same time each variable presents something unique.

The variable _rss_score provides information about the quantity of different RSS feeds found in a single website in relation to the total pages crawled for that website. As seen in the variable's histogram in Figure A1a the relevant majority of websites do not provide RSS feeds at all. That being said, the sites that do provide RSS in total are more than the sites that do not. From the websites that do use RSS as a means to disseminate content most have an _rss_score value from 1–100. This indicates that they provide one or less unique RSS feed per page crawled. With the arithmetic mean being ~20, as seen in Table 5, that would mean that the average website provided 1 RSS feed per 5 pages crawled. This makes sense since usually RSS feeds contain multiple records of content (articles, products, comments, etc.) It was observed that a common practice was to provide some general feeds with multiple records while at the same time providing an additional single record feed in a single article or artwork page. Very few websites seem to provide an abnormally large number of RSS feeds. The vast majority of these websites are sites with very few pages crawled (one or two), which included multiple feeds. These cases account for less than 2% of total websites and might be the result of technical irregularities. In general, RSS usage seems to remain somewhat popular despite the technology being past its prime. A contributor to this might be the fact that many popular Content Management Systems (such as WordPress or Wix) provide RSS feed support out-of-the-box.

The histogram of the variable _html_score in Figure A2a represents a duality: A large number of websites use the HTML Semantic Element tags in all of their pages and another smaller number in none. This is to be expected since adoption of such a technology often happens in a vertical manner throughout all the pages of a website. The mean of the variable is ~77, indicating that usage of at least some HTML Semantic Elements is rather popular. We can obtain more insights regarding these elements through the _html_variety variable. Its histogram in Figure A6a, shows a relatively normal distribution. The peak of

that distribution is at 46% which means that most websites that use these elements use 6 out of 13 identified elements. Although the variety of elements used could be higher, the overall assessment of HTML5 Semantic Elements is encouraging since the technology is both popular and focused in more than a few different elements. Further observation in this area can provide information on the more or least used such elements but it is beyond the scope of this study.

The social media related variables _og_score and _twitter_score have similar histograms, as seen in Figures A3a and A4a, with the bulk of websites either fully committing to the technology or not implementing it at all. This behavior that was also noted in _html_score seems to form a pattern. Open Graph seems to be the more popular of the two with a mean of ~65 vs. one of ~31 as seen in Table 5. This is to be expected since Open Graph is used by multiple social media platforms and messaging systems to create rich previews. Even Twitter itself will create a rich object through open graph if no Twitter card is available. A Twitter card can even indicate that a preview's content can be collected by the appropriate Open Graph meta elements. Looking at the _og_variety and _twitter_variety variables in Figures A7a and A8a, we can note that most websites that implement the technologies also ensure that they provide unique information for each different page of the website. This builds into the already established pattern that when a website developer decides to implement such a technique the implementation is usually comprehensive. Although fewer, there are still cases of websites that provided non-unique titles for the rich object preview.

The structured data related variable _schema_score showed a moderate usage of the schema.org vocabularies throughout the websites included in this study as it was indicated by its mean which is at ~42 as seen in Table 5. In its histogram in Figure A5a we notice the same behavioral pattern as other similar variables. In contrast, the other variables used to identify structured data usage (_microformats, _microformats_variety, _other) all recorded very low usage around 1%. A secondary crawling trying to identify elements with Microformats v2 classes yielded even fewer websites. This indicates that website developer efforts towards implementing structured data is for the time being focusing mainly on schema.org which is founded and supported by major players in the field of SEO and the Web in general.

The Semantic Web Technologies Integration rating as described in Section 3.3 tries to summarize all above metrics in an overall rating (variable _stwi). This variable had a mean of ~2.9 as seen in Table 5 which indicates and above average integration and a standard deviation of ~1.5. In Table 6 the percentile breaking points at 25%, 50% and 75% for this variable are 1.5, 3 and 4.5 which can be interpreted as an indicator of the rating system's quality. In the histogram of the variable, as seen in Figure A9a, we notice two peaks, one around rating value 2 and one around rating value 4.5. This double peak impression can be a result of the behavioral pattern of either implementing a technology fully or not at all that we discerned in the histograms of other individual variables.

As described in Section 3.1.2, the websites were screened by the researchers and split into categories: Red websites, that were outside the study's scope and were not crawled, Yellow, which indicated that a website was art or culture related but had limited reportorial content and Green, which indicated that a website was not only art or culture related but also contained a fair amount of reportorial content. In Section 4 we proceeded to distinguish the information between these two classes, thus creating the variables _swti_yellow and _swti_green. We can see from Table 7 that the SWTI for Green websites has a mean of ~3.2 which is not only greater than that of the yellow websites but also greater than the overall mean of _swti. Additionally, in the histograms of these new variables we notice an overall shift of frequency values towards higher STWI ratings. This is a fair indicator that websites that purposefully provide more journalistic or reportorial contact concerning art and cultural heritage also put more effort into implementing Semantic Web technologies.

Studying the interrelationship between several of the variables that were calculated using metrics from the crawling algorithm described in Section 3.2, there appear to be

multiple moderate positive correlations between them, as seen in Table 9. This is a clear indication that when developers decide to start integrating Semantic Web technologies in their websites, they will often branch to multiple such technologies in order to achieve more comprehensive coverage. The strong correlation between _og_score and _twitter_score is also notable since it demonstrates the importance of multiple technologies when focusing on social media. Developers do not always go for one technology over the other, but they display a notable preference to implement both.

The Spearman correlation analysis between the ranking of the websites based on their SWTI rating and their Alexa ranking indicates a very small, yet significant positive correlation which means that to a small extent, usage of Semantic Web technologies and website popularity are indeed positively related. This can indicate both that websites that are popular are more keen to invest in Semantic Web integration and that Semantic Web integration might actually provide a minimal boost to popularity. To make more of this linear relationship the researchers proceeded to a gradient boosting analysis the results of which were presented in Section 4.2.

The Gradient Boosting analysis was performed as mentioned using the XGBoost algorithm and provided some interesting findings. We can see from the overall statistics presented in Table 11 that overall prediction accuracy surpassed the value of 0.3. Considering the random prediction accuracy for the four defined intervals would be 0.25 there appears to be a small but noticeable increase. The increase's persistence can be observed by the minimum and maximum values of the 95% confidence interval which are both above the baseline of 0.25. Moreover, this increase indicates that, even though Semantic Web integration as measured by this study is not directly correlated with each website's overall popularity, it can still be used to an extent to more accurately predict under which popularity class a website would fall.

By assessing the statistics by class as seen in Table 12, it appears that Class 1, which includes the top 25% of the websites in ranking, displays higher values than the other classes in Positive Prediction Value, Precision and Balanced Accuracy. This might indicate higher credibility of Semantic Web metrics when attempting to predict the popularity of top-ranking websites.

In Figure 8 the features used in the Gradient Boost analysis are presented by order of calculated importance in the accurate prediction of popularity. They are clustered in four groups according to that importance. First and only feature in the most important cluster is the "_schema_org" feature which is an indicator of the percentage of crawled pages that include schema.org structured data. Usage of the schema.org vocabularies is promoted by Google, Microsoft, Yahoo, Yandex and more search engine providers which means that their inclusion to a larger extent, not only provides machine-readable content, but also increases the website's Search Engine Optimization Score (SEO) which in turn influences popularity.

In the second cluster the features _og, _twitter and _html_variety appear. The first two assist with social media integration and thus make a page easier to diffuse through the multiple social media platforms available. The _html_variety feature represents the effort, from a developer's perspective, to enhance a web page's semantic value by using a greater variety of HTML5 semantic elements.

The other features appear to have less importance and are grouped in the remaining clusters. Social media-related rich-data content variety as indicated by "_og_variety" and "_twitter_variety" seems to matter but to a smaller extent. This makes sense if we consider that content can sometimes be accurately described even without much variation in the description itself. The feature "_rss_feeds" which indicates the usage of RSS also plays a more minor role. RSS, though still useful, seems to be of waning importance as a means to convey machine readable information. Additionally, all metrics relating to microformats appear to be irrelevant. This is to be expected judging by how few implementations of this Semantic Web tool were detected during the crawling process.

6. Conclusions

The Semantic Web since its conception has been embraced by people in the fields of art and cultural heritage, because it provides valuable tools for organizing and disseminating digital data, regarding not only works of art and culture but also content relevant to these works, such as reportorial and academic articles, reviews, opinion pieces, event or exhibition retrospectives, and more. This study has shed a light on the level of integration of Semantic Web technologies and how the various metrics that quantify different Semantic Web technologies can be used not only to assess Semantic Web integration, but might also influence or predict website popularity to a small extent.

According to the findings, many of the distributions of the various variables displayed a pattern of having two peaks, one at the lowest and one at the highest value. This indicates that most websites either completely ignore the use of a specific Semantic Web technology or fully commit to it, implementing it comprehensively. Additionally, the moderate correlations between the various metrics indicated that integration with the Semantic Web as a general goal is mostly either ignored or pursued thoroughly. Finally, through the Gradient Boosting analysis it was established that the integration of schema.org structure data in a website was the most important factor in the ability to predict the website's popularity.

The research presented in this study was limited in its ability to fully include all relevant websites. Additional insight might be found in websites that might not have been discovered by the study's methodology or that were excluded for not being in English. Further research, monitoring Semantic Web integration in the websites of developing countries such as China or India might produce different results and assist in creating a more comprehensive overview of the landscape of Semantic Web technologies integration. Additionally, the present research focused exclusively in the areas of art and culture, but things might be different in other fields. The line of research presented here can continue in the future, with the focus shifting from media relating to art and culture to media relating to other fields such as sports, technology, consumer products, and more. The Semantic Web Technologies Integration rating introduced is content-agnostic and as such can be used to evaluate integration in any field. Additionally, its simplicity allows its use even without the automated crawling algorithm described in this article, as long as the data set of relevant websites is small. Enriching the data-gathering process with even more technologies that encompass aspects of the Semantic Web as they become popular in the future is also important and can form a basis for future research.

Studying and analyzing the tangible presence of the Semantic Web is an important step in evaluating its progress and can be of valuable help in achieving its true potential, which so far remains largely untapped. The increased relevance of social media and the marketing importance of SEO can both become incentives to further expand both the quantity and the quality of machine-readable structured rich data in websites of any magnitude or topic through technologies such as Open Graph and Schema.org. Furthermore, new challenges emerge with the decentralization principles brought forward with the popularization of blockchain technology and the Semantic Web must rise to meet them in order to expand and encompass all aspects of the World Wide Web as it evolves with unprecedented celerity.

Author Contributions: Conceptualization, A.G.; Formal analysis, M.P., N.K. and A.K.; Investigation, A.G., M.P. and A.L.; Methodology, A.G. and M.P.; Project administration, A.G.; Resources, N.K. and A.K.; Software, M.P.; Supervision, A.G. and I.V.; Visualization, A.L.; Writing—original draft, M.P., N.K. and A.K.; Writing—review and editing, M.P. and I.V. All authors have read and agreed to the published version of the manuscript.

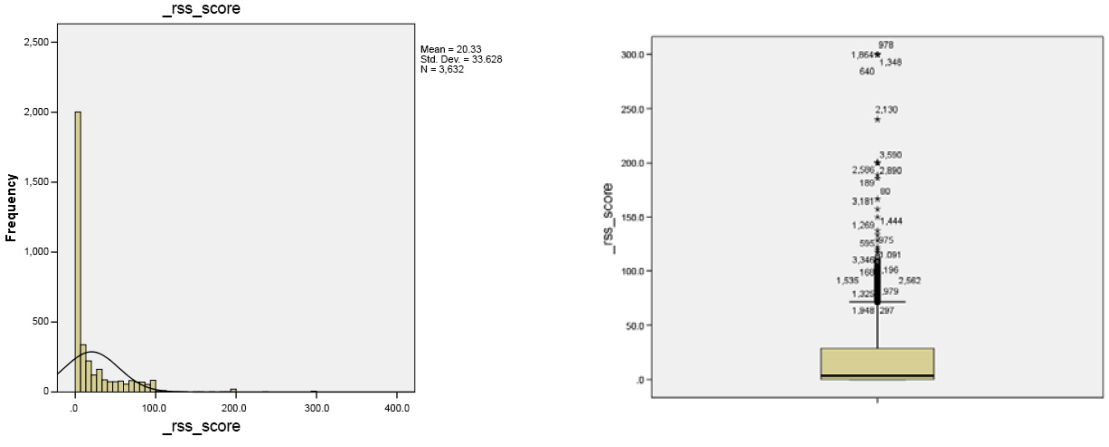
Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Zenodo at [10.5281/zenodo.5811988], reference number [10.5281/zenodo.5811988].

Conflicts of Interest: The authors declare no conflict of interest.

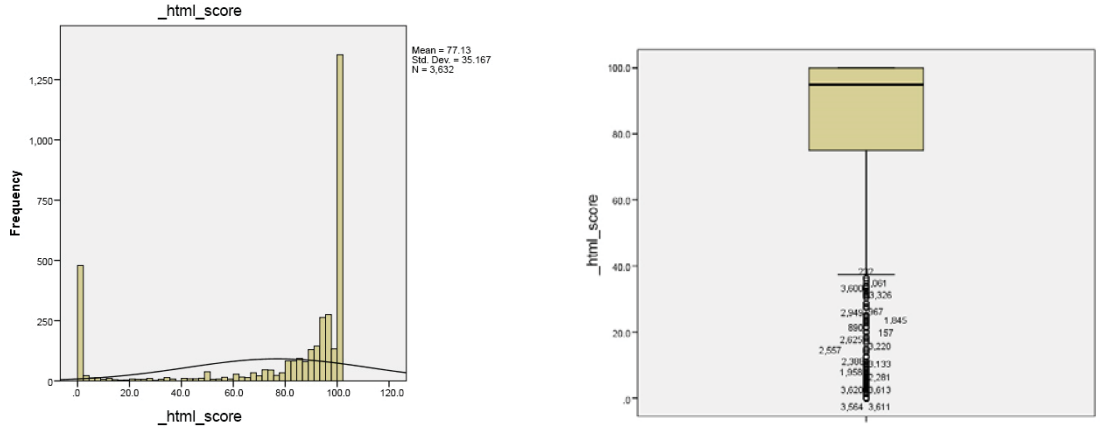
Appendix A

This appendix presents the histogram and boxplots for the variables _rss_score, _html_score, _og_score, _twitter_score, _schema_score, _html_variety, _og_variety, _twitter_variety and _swti. These distributions and dispersions are discussed in detail in Section 5 of this article.



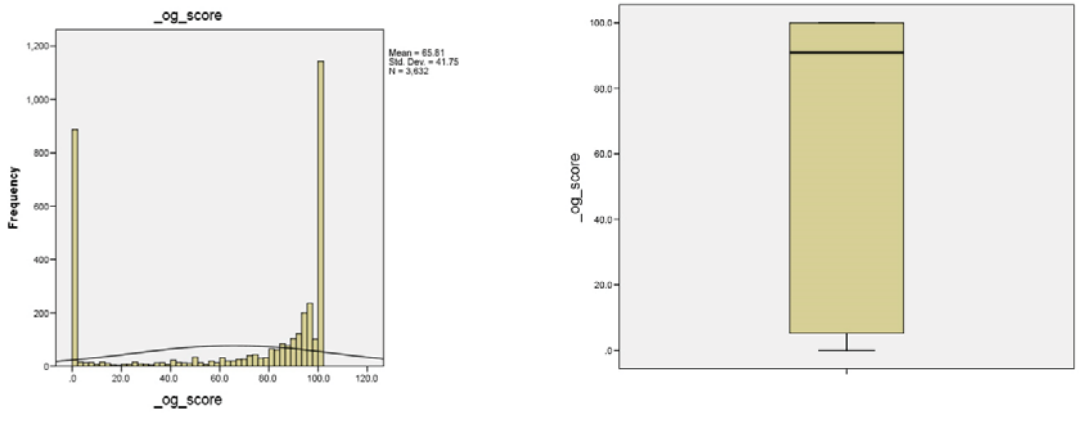
(**a**) (**b**)

Figure A1. The histogram (**a**) and boxplot (**b**) of the _rss_score variable showing a non-normal distribution.



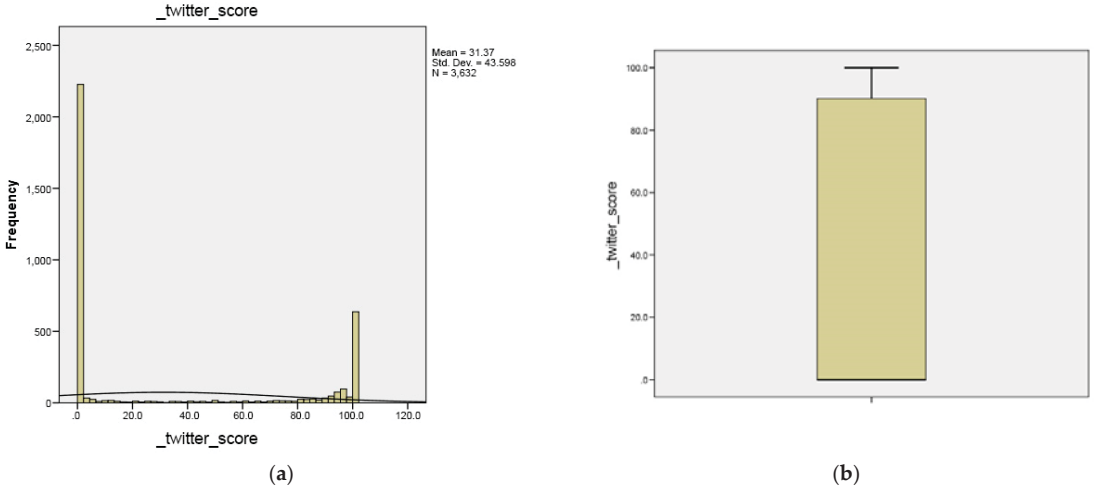
(**a**) (**b**)

Figure A2. The histogram (**a**) and boxplot (**b**) of the _html_score variable showing a non-normal distribution.



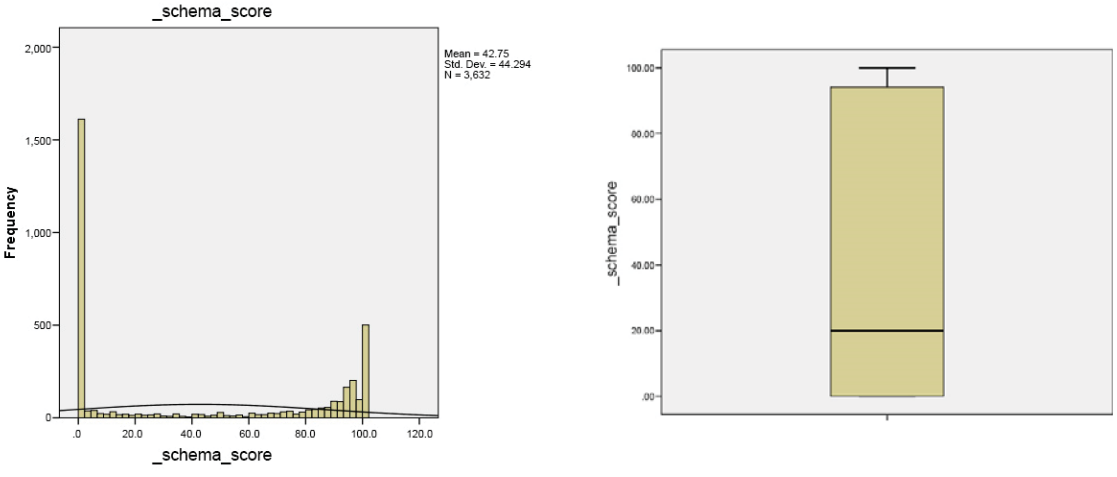
(**a**) (**b**)

Figure A3. The histogram (**a**) and boxplot (**b**) of the _og_score variable showing a non-normal distribution.



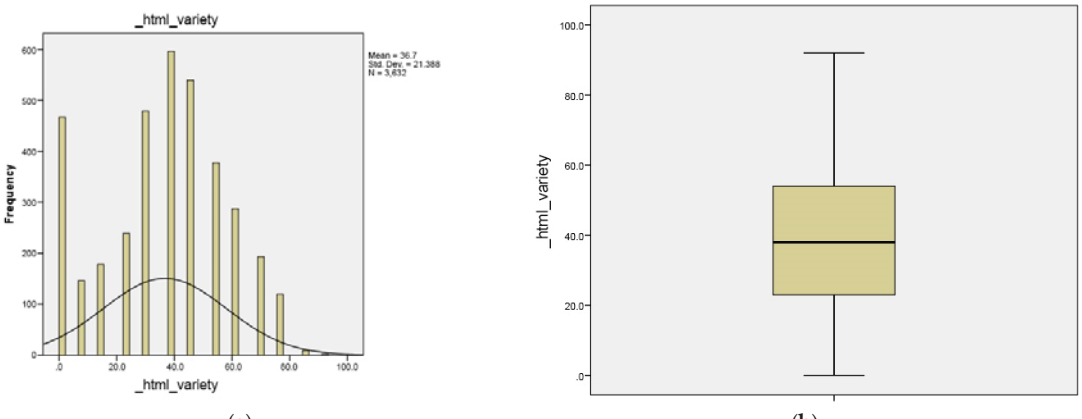
(**a**) (**b**)

Figure A4. The histogram (**a**) and boxplot (**b**) of the _twitter_score variable showing a non-normal distribution.



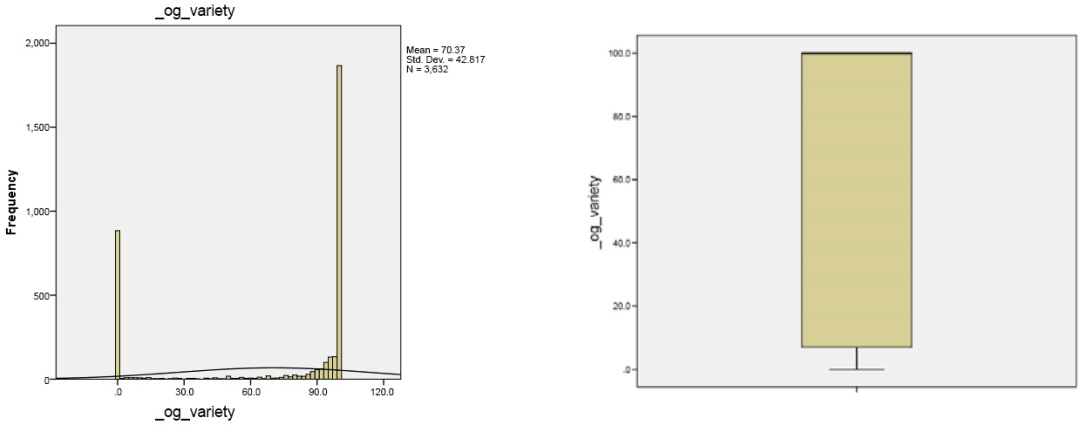
(**a**) (**b**)

Figure A5. The histogram (**a**) and boxplot (**b**) of the _schema_score variable showing a non-normal distribution.



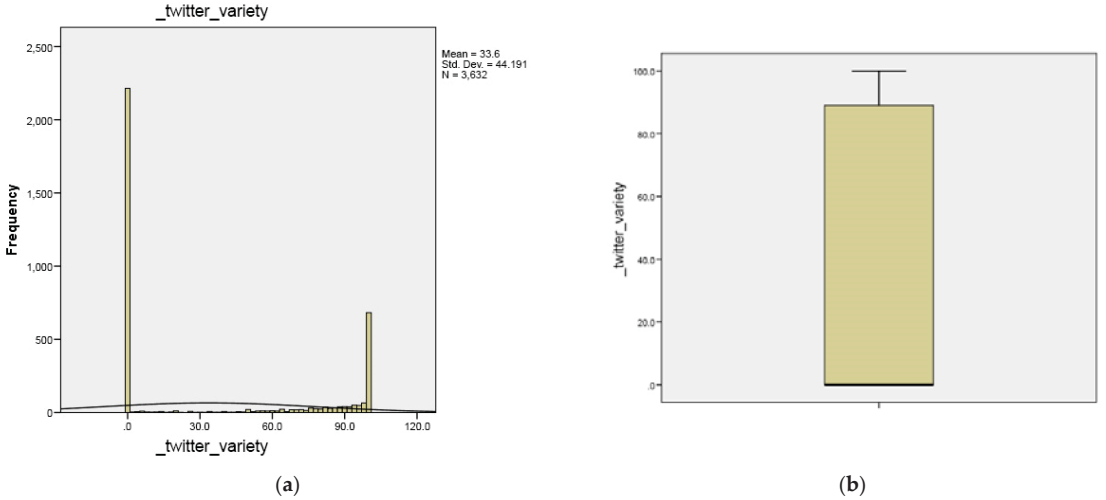
(**a**) (**b**)

Figure A6. The histogram (**a**) and boxplot (**b**) of the _html_variety variable showing a non-normal distribution.



(**a**) (**b**)

Figure A7. The histogram (**a**) and boxplot (**b**) of the _og_variety variable showing a non-normal distribution.



(**a**) (**b**)

Figure A8. The histogram (**a**) and boxplot (**b**) of the _twitter_variety variable showing a non-normal distribution.

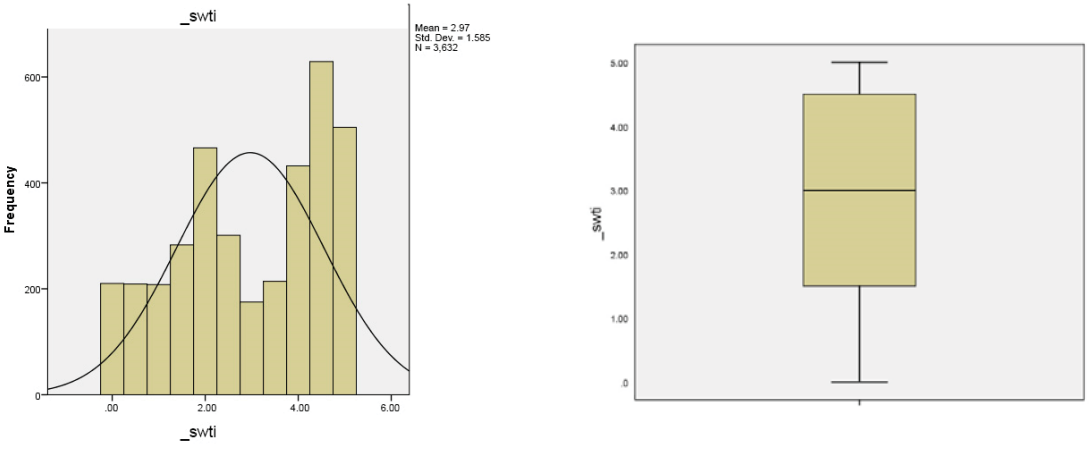


Figure A9. The histogram (**a**) and boxplot (**b**) of the _swti variable showing a non-normal distribution.

References

1. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
2. Bing, L.; Chan, K.C.; Carr, L. Using aligned ontology model to convert cultural heritage resources into semantic web. In Proceedings of the 2014 IEEE International Conference on Semantic Computing, Newport Beach, CA, USA, 16–18 June 2014; pp. 120–123.
3. Panagiotidis, K.; Veglis, A. Transitions in Journalism—Toward a Semantic-Oriented Technological Framework. *J. Media* **2020**, *1*, 1–17. [CrossRef]
4. Heravi, B.R.; McGinnis, J. Introducing social semantic journalism. *J. Media Innov.* **2015**, *2*, 131–140. [CrossRef]
5. Lim, Y.S. Semantic web and contextual information: Semantic network analysis of online journalistic texts. In *Recent Trends and Developments in Social Software*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 52–62.
6. Dertouzos, M. *The Unfinished Revolution: Human-Centered Computers and What They Can Do for Us*; HarperCollins: New York, NY, USA, 2001.
7. Nath, K.; Dhar, S.; Basishtha, S. Web 1.0 to Web 3.0—Evolution of the Web and its various challenges. In Proceedings of the ICROIT 2014—2014 International Conference on Reliability, Optimization and Information Technology, Faridabad, India, 6–8 February 2014; pp. 86–89. [CrossRef]
8. Varlamis, I.; Giannakoulopoulos, A.; Gouscos, D. Increased Content Accessibility For Wikis And Blogs. In Proceedings of the 4th Mediterranean Conference on Information Systems MCIS, Athen, Greece, 25–27 September 2009.
9. Belk, M.; Germanakos, P.; Tsianos, N.; Lekkas, Z.; Mourlas, C.; Samaras, G. Adapting Generic Web Structures with Semantic Web Technologies: A Cognitive Approach. In Proceedings of the 4th International Workshop on Personalised Access, Profile Management, and Context Awareness in Databases (PersDB 2010). 2010. Available online: http://www.vldb.org/archives/website/2010/proceedings/files/vldb_2010_workshop/PersDB_2010/resources/PersDB2010_6.pdf (accessed on 30 December 2021).
10. Blake, J. On defining the cultural heritage. *Int. Comp. Law Q.* **2000**, *49*, 61–85. [CrossRef]
11. Kabassi, K. Evaluating museum websites using a combination of decision-making theories. *J. Herit. Tour.* **2019**, *14*, 1–17. [CrossRef]
12. Kiourexidou, M.; Antonopoulos, N.; Kiourexidou, E.; Piagkou, M.; Kotsakis, R.; Natsis, K. Multimodal Technologies and Interaction Websites with Multimedia Content: A Heuristic Evaluation of the Medical/Anatomical Museums. *Multimodal Technol. Interact.* **2019**, *3*, 42. [CrossRef]
13. Dannélls, D.; Damova, M.; Enache, R.; Chechev, M. A framework for improved access to museum databases in the semantic web. In Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, 16 September 2011; pp. 3–10.
14. Dimoulas, C.; Veglis, A.; Kalliris, G. Audiovisual Hypermedia in the Semantic Web. In *Encyclopedia of Information Science and Technology*, 3rd ed.; IGI Global: Hershey, PA, USA, 2015; pp. 7594–7604. [CrossRef]
15. Janssen, S. Art journalism and cultural change: The coverage of the arts in Dutch newspapers 1965–1990. *Poetics* **1999**, *26*, 329–348. [CrossRef]
16. Matthews, B. Semantic web technologies. *E-learning* **2005**, *6*, 8.
17. Kumar, S.; Chaudhary, N. A Novel Trust Scheme in Semantic Web. In *Information and Communication Technology for Sustainable Development*; Springer: Singapore, 2020; pp. 103–110.
18. World Wide Web Concortium (W3C). Available online: https://www.w3.org/ (accessed on 20 December 2021).

19. Ning, X.; Jin, H.; Wu, H. RSS: A framework enabling ranked search on the semantic web. *Inf. Process. Manag.* **2008**, *44*, 893–909. [CrossRef]
20. Necula, S.C.; Păvăloaia, V.D.; Strîmbei, C.; Dospinescu, O. Enhancement of e-commerce websites with semantic web technologies. *Sustainability* **2018**, *10*, 1955. [CrossRef]
21. Patel-Schneider, P.F. Analyzing schema. org. In *International Semantic Web Conference*; Springer: Cham, Switzerland, 2014; pp. 261–276.
22. Meusel, R.; Bizer, C.; Paulheim, H. A web-scale study of the adoption and evolution of the schema. org vocabulary over time. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, Larnaca, Cyprus, 13–15 July 2015; pp. 1–11.
23. Mika, P. On schema. org and why it matters for the web. *IEEE Internet Comput.* **2015**, *19*, 52–55. [CrossRef]
24. Common Crawl. Available online: https://commoncrawl.org/ (accessed on 20 December 2021).
25. Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Lamprogeorgos, A.; Limniati, L.; Varlamis, I. Exploring the Dominance of the English Language on the Websites of EU Countries. *Future Internet* **2020**, *12*, 76. [CrossRef]
26. Alonso, I.; Bea, E. A tentative model to measure city brands on the Internet. *Place Branding Public Dipl.* **2012**, *8*, 311–328. [CrossRef]
27. Govers, R.; Van Wijk, J.; Go, F. Website Analysis: Brand Africa. In *International Place Branding Yearbook 2010: Place Branding in the New Age of Innovation*; Palgrave Macmillan: London, UK, 2010; pp. 156–171. [CrossRef]
28. Alexa Internet-About Us. Available online: https://www.alexa.com/about (accessed on 20 December 2021).
29. PHP cURL Introduction. Available online: https://www.php.net/manual/en/intro.curl.php (accessed on 20 December 2021).
30. Powers, S. *Practical RDF*; O'Reilly Media, Inc.: Cambridge, MA, USA, 2003; p. 10.
31. HTML Semantic Elements. Available online: https://www.w3schools.com/html/html5_semantic_elements.asp (accessed on 20 December 2021).
32. The Open Graph Protocol. Available online: https://ogp.me/ (accessed on 20 December 2021).
33. About Twitter Cards. Available online: https://developer.twitter.com/en/docs/twitter-for-websites/cards/overview/abouts-cards (accessed on 20 December 2021).
34. Microformats—Building Blocks for Data-Rich Web Pages. Available online: https://microformats.org/ (accessed on 20 December 2021).
35. Roussos, P.L.; Tsaousis, G. *Statistics in Behavioural Sciences Using SPSS*; TOPOS: Athens, Greece, 2011.
36. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
37. Zhang, L.; Zhan, C. Machine learning in rock facies classification: An application of XGBoost. In Proceedings of the International Geophysical Conference, Qingdao, China, 17–20 April 2017; Society of Exploration Geophysicists and Chinese Petroleum Society: Tulsa, OK, USA, 2017; pp. 1371–1374.
38. Steinberg, D. Classification and regression trees. In *The Top Ten Algorithms in Data Mining*; Wu, X., Kumar, V., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009; pp. 179–202.
39. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
40. Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 161–168.
41. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]

future internet

MDPI

Article

A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing

Nikolaos Vryzas *, Anastasia Katsaounidou, Lazaros Vrysis , Rigas Kotsakis and Charalampos Dimoulas

Multidisciplinary Media & Mediated Communication Research Group (M3C), Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; akatsaounidou@gmail.com (A.K.); lvrysis@auth.gr (L.V.); rkotsakis@auth.gr (R.K.); babis@auth.gr (C.D.)

* Correspondence: nvryzas@auth.gr

Abstract: Media authentication relies on the detection of inconsistencies that may indicate malicious editing in audio and video files. Traditionally, authentication processes are performed by forensics professionals using dedicated tools. There is rich research on the automation of this procedure, but the results do not yet guarantee the feasibility of providing automated tools. In the current approach, a computer-supported toolbox is presented, providing online functionality for assisting technically inexperienced users (journalists or the public) to investigate visually the consistency of audio streams. Several algorithms based on previous research have been incorporated on the backend of the proposed system, including a novel CNN model that performs a Signal-to-Reverberation-Ratio (SRR) estimation with a mean square error of 2.9%. The user can access the web application online through a web browser. After providing an audio/video file or a YouTube link, the application returns as output a set of interactive visualizations that can allow the user to investigate the authenticity of the file. The visualizations are generated based on the outcomes of Digital Signal Processing and Machine Learning models. The files are stored in a database, along with their analysis results and annotation. Following a crowdsourcing methodology, users are allowed to contribute by annotating files from the dataset concerning their authenticity. The evaluation version of the web application is publicly available online.

Keywords: tampering; authentication; misinformation; web application; news; machine learning; deep learning; crowdsourcing

Citation: Vryzas, N.; Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing. *Future Internet* **2022**, *14*, 75. https://doi.org/10.3390/fi14030075

Academic Editor: Carlos Filipe Da Silva Portela

Received: 7 January 2022
Accepted: 24 February 2022
Published: 27 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

News authentication is considered a vital task for reliable informational services. The COVID-19 pandemic situation that we currently experience showcased the importance of fact-checking in fighting disinformation to protect our societies and democracies. The role of audiovisual recording is considered crucial in documenting news articles, thus convincing audiences about the truth of the underlying events [1–3]. With the advancement of Information and Communication Technologies and the availability of easy-to-use editing and processing tools, one unwanted side-effect is the falsification of multimedia assets (i.e., images, audio, video) to alter the presented stories, making them more appealing (or intentionally doctored). In this context, unimodal solutions have been implemented to inspect each of the individual media entities, while multimodal forensic services are also deployed through online collaborative environments, plug-ins, serious games, and gamification components [1,2,4,5].

While the detection of manipulated photos/images and the evaluation of the associated forgery attacks remain critical [6], audio and video content have become even more popular nowadays. In this context, audio offers some unique features, such as less demanding processing needs and the inherent time continuity, making tampering inconsistencies easier to reveal [7,8]. Semantic processing and machine learning technologies empower

today's digital forensics tools. However, these new capabilities can also be exploited for counter-/anti-forensic means, requiring constant and continuous effort.

1.1. Related Work

Content verification has always been a very important part of journalistic workflows and a crucial factor of journalistic ethics and deontology. In the context of Journalism 3.0, where new business models of low or no pay journalism, combined with news aggregation and republishing and the reuse of amateur user-generated content (UGC) [9], disinformation has become a major problem for journalistic practice. As a result, several fact-checking organizations have appeared in the past decade, intending to find and debunk false claims that are spread throughout the Web and social media services [10]. Recent research of academics and organizations has been directed towards highlighting the best practices for content verification, through international cooperation networks [11].

In the modern media ecosystem, data variety is a very important parameter of big data volumes [12]. This means that fact-checkers need to manage content in many different modalities (e.g., text, audio, image, video). Different approaches and methodologies have to be defined for each case [13]. In disinformation, media assets may be used in a misleading context to support a false claim, or may be manipulated themselves. In the first case, an image/audio/video file is followed by an untrue description or conclusion, while in the latter, the media file has been maliciously edited. Such manipulations may include actions, such as copying and moving parts of the file to a different place and splicing in segments of a different file, aiming at affecting the semantic meaning of the file [14]. Common cases can be found in all file types, whether image, audio, or video. In the case of image tampering detection, spatial techniques can be used to locate suspicious regions and discontinuities within an image file. Media Verification Assistant is a project that allows users to upload images and applies several algorithms to provide forensics analysis [14,15]. In contrast to static images, audio and video files introduce the dimension of time. In video files, besides the spatial analysis of single image frames, the detection of temporal discontinuities can be crucial for the spatiotemporal location of malicious tampering [16]. Such techniques are expected to be computationally heavy. Audio is a very important modality present in the majority of video files. In this sense, audio can be used autonomously for the authentication of both audio and video assets. Audio information retrieval techniques are much less computationally complex. Audio forensics tools are not, however, as well-explored as those applied to visual information. Two important toolboxes on the market are the ARGO-FAAS [17] and the EdiTracker plugin [1]. They are, however, paid services, and not publicly available.

Audio forensics techniques address the processes of audio enhancement, restoration, and authentication of an audio asset so that it can be considered as evidence in court [18,19]. Authentication techniques aim at detecting artifacts within an audio file that can indicate malicious editing. Traceable edits can be found in the file container information or in the audio content [20,21]. Techniques that inspect file container inconsistencies investigate the metadata, descriptors, or the encoding structure. When the audio content is investigated, the aim is to use dedicated software to detect certain artifacts that may be inaudible by human subjects. Several different approaches can be found in the literature.

Electronic Network Frequency (ENF) techniques make use of the phenomenon of the unintentional recording of an ENF through interference. Electronic networks provide alternating current with a nominal frequency of 50 or 60 Hz, depending on the region. However, the real frequency of the current fluctuates around this value. The electronic equipment that is used for recordings captures this frequency fluctuation, which can act as a timestamp of the recording. It is possible to isolate and track the ENF in recordings to check whether there is phase inconsistency in the fluctuation, or even to find the exact time of the recording from the log files of the electronic networks [22–25].

Other approaches investigate the acoustic environment of the recording, such as the Signal-to-Reverberation ratio of a room [26,27]. The specifications of a recording device

have also been proven to be traceable in research, providing an indicator of whether parts of an audio file were recorded with a different device [20,28–30]. Dynamic Acoustic Environment Identification (AEI) may rely on statistical techniques that rely on reverberation and background noise variance in a recording [31]. Machine learning techniques are proven to be very useful for acoustic environment identification. Machine learning models do not rely on the definition of a set of rules for decision-making, but require a dataset of pre-annotated samples to train a classification model that can identify different classes, in this case, acoustic environments [31–33].

Another methodology for audio tampering detection investigates file encoding and compression characteristics. A huge number of highly configurable audio encodings are available that differ in terms of compression ratio, bitrate, use of low-pass filters, and more. A file that comes from audio splicing is very likely to contain segments encoded with different configurations, which can be traceable [34–36]. Most encoding schemes depend on psychoacoustic models that apply algorithms to discard redundant, inaudible frequencies. The Modified Discrete Cosine Transform (MDCT) coefficients can be investigated using statistical or machine learning methods to detect outliers in specific segments of the file [37]. Even when the file is reencoded in another format, there are often traces of the effect of previous compression algorithms [38–41].

Media authentication can be supported during content production using container and watermarking techniques, such as hash-code generation and encryption, and MAC timestamp embedding. Recovery of the inserted hash code that was generated by algorithms, such as SHA-512, enables the detection of tampered points within an audio stream [42]. Similarly, embedding timestamp information in files can allow the identification of an audio excerpt with a different MAC timestamp that has been maliciously inserted [20].

Whether the aim is training machine learning models or evaluating proposed analysis methods, one crucial part of every audio authentication project is the formation of a dataset. This is a very complex procedure due to the task's peculiarities, and it often acts as a bottleneck for the robustness of such techniques. Not many datasets are available for experimentation. In [43], a dataset was recorded featuring different speakers, acoustic rooms, and recording devices. In [44] a dataset with different encodings was created through an automated process. In [45], existing recordings were edited to create a dataset. In [7], an automated process was proposed for the creation of a multi-purpose dataset using an initial set of source files provided by the user.

1.2. Project Motivation and Research Objectives

It has been made clear that machine learning solutions for audio tampering detection require a dataset for the training of models. Since datasets with real cases of tampered files are not available, most works require the formulation of artificial datasets for model evaluation. Such datasets are often difficult to handcraft, so they follow automated procedures for dataset creation, simulating real-world scenarios. As a result, the implemented models are case- and dataset-specific. There is no evidence for the generalization of the models in multiple scenarios and tampering techniques. For this reason, it is not yet feasible to integrate automated audio authentication into professional workflows without supervision, as they cannot be considered reliable for production and real-world applications. Furthermore, models that are pre-trained in known datasets and conditions may be more vulnerable to adversarial attacks [46].

On the other hand, traditional audio forensics techniques require expertise and fluency with audio analysis tools. In such an approach, human intelligence and experience play a crucial role in the process of authentication. While this is the most reliable solution and the preferable option in courtrooms, it cannot provide a viable alternative with massive appeal. There is an urgent need for tools that can help in the fight against disinformation. Such tools should be accessible to a broad audience of journalists, content creators, and simple users, to improve the overall quality of news reporting. Average users do not have the expertise to apply audio analysis techniques in the same way as professionals of audio forensics.

The motivation for the current research emerges from the hypothesis that it is feasible to strengthen a user's ability to recognize tampered multimedia content using a toolbox of supervisory tools provided online through an easy-to-use interface. State-of-the-art approaches for audio analysis and tampering detection were integrated into a web application. The application is available publicly through a web browser. The results of the algorithms do not provide an automated decision-making scheme, but rather a set of visualizations that can assist the user in a semi-automated approach. This means that the framework does not include a model that performs binary classification of files as tampered, or not-tampered, but the final decision is the responsibility of the user, taking advantage of their perception and experience as well as the context of the media asset. Through the use of the application, crowdsourcing is promoted for the creation of a dataset with real-world tampered files for future use.

The remaining of the paper is structured as follows. In Section 2, the proposed web framework is presented, in terms of the functionality, aims, and technical specifications. The integrated algorithms and their operating principles are listed without emphasizing technical details. In Section 3, the evaluation results from the reverberation estimation models and the initial implementation of the prototype web application are presented. In Section 4, the research results are summarized and discussed, and the future research goals of the project are defined. In Section 5, some of the limitations of the presented research are analyzed.

2. Materials and Methods

As stated in the problem definition section, the proposed approach consists of a framework for the assistance of professional journalists and the public in detecting tampered audiovisual content. The core of the framework is a web application with a graphic user interface provided to the public for the submission and analysis of content. The application incorporates an ensemble of algorithms that provide the user with supervisory tools for semi-automatic decision-making. The analysis strategy is audio-driven, as it makes use of the audio channel. The integrated algorithms do not classify files as tampered or not, but rather support the users in decision-making. The application offers the necessary crowdsourcing functionality for dataset creation and user cooperation. The framework was designed and implemented as a component of the Media Authentication Education (MAthE) project, which aims at providing educational and gamification tools to battle misinformation [4].

2.1. A Web Application for Audio Tampering Detection and Crowdsourcing

The main goal of the web application is to combine the effectiveness of state-of-the-art signal processing, machine learning advances and human perception for computer-assisted audio authentication. The application:

1. Implements state-of-the-art analysis options. An ensemble of algorithms is incorporated, addressing multiple audio tampering strategies. Such strategies may include encoding detection, recording conditions, background noise clustering, and others.
2. Follows a modular approach. The algorithms that are provided in the initial implementation are available as individual modules. This allows the existing algorithms to be upgraded in the future, as well as the extension of the initially provided toolbox.
3. Supports human-centered decision-making. As was explained, it is within the rationale of the MaThe solutions to promote computer-assisted decision making. The algorithmic implementations provide intuitive visualizations aiming at assisting the user in content authentication, taking also into consideration the user's personal experience and perception, as well as the context of the asset under investigation.
4. Is publicly available. As was explained, the web framework aims to address a wide public. An important prerequisite for this is that it is freely available for anyone to use and contribute.

5. Requires no audio or technical expertise. The design principles prioritize ease-of-use, following a typical workflow. A more experienced user with a technical and signal processing background, may get better insight and understanding of the produced visualizations. However, the detection of outliers or suspicious points in a file timeline is self-explanatory and does not require a deep understanding of the algorithms and mechanisms.
6. Promotes crowdsourcing. Users and teams can become involved and contribute to the project in several ways to further advance the field of audio tampering detection. They can submit files, annotated as tampered or not tampered, with a brief justification. Users can also randomly browse files from the dataset, analyze them, and mark them as tampered or not tampered. Finally, as this is an open-source project following a modular architecture, researchers and teams are encouraged to contribute with code and extensions.

The main functionality of the MAthE AudioVisual Authentication framework is shown in Figure 1. Users can submit files for analysis and investigation, or contribute by annotating existing files concerning their authenticity. Once a file is submitted, the application returns analysis results, and the user can decide if they want to submit the file to the database along with an annotation (tampered or not tampered), submit the file to the database without annotation, or not submit anything to the database. Contributing users can access submitted files, annotated or not, examine the analysis results, and provide annotation (tampered or not tampered), following a crowdsourcing methodology.

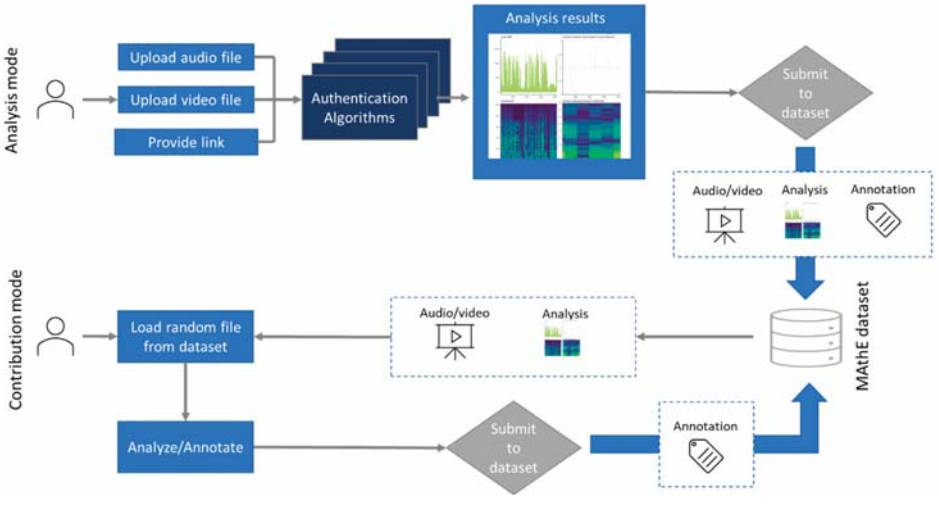


Figure 1. The MAthE AudioVisual Authentication framework and functionality.

2.2. The Computer-Supported Human-Centered Approach

The main concept of our approach depends on the idea that actors with no expertise in signal processing, machine learning, and computational methods can benefit from the visualization output of such techniques with little or no training. Computational methods in media authentication usually try to detect anomalies within the file under investigation. Such anomalies can be visually depicted (e.g., with a change in color). A non-expert user can perceive such depictions and interpret them accordingly, even without understanding or knowledge of the technical details that led to this visualization. After locating the suspicious points within the file, the user can base their reaction based on contextual information and their own critical thought. For example, an object within an image that looks like an anomaly in the visualizations and also dramatically alters the

image's semantic meaning probably indicates tampering. This is the main idea of the ReVeal project [14], which deals with tampered images and was a major inspiration of the present work. There is evidence from experiments that users with no technical knowledge were able to detect tampering of images with the support of such visualizations [6]. In this approach, a gamification approach was also tested that allowed users to ask for the help of such a visualization toolbox [5] in order to detect fake news and proceed in the game [4]. While such techniques are not robust in the automated detection of media content tampering, they can push the limits of human intellect and support users to make better decisions on fake content recognition.

In this direction, in the present work, several visualizations based on anomalies that are detected by audio processing are proposed. Since the audio channel is commonly part of video files, this toolbox aims at supporting users with no technical expertise to make decisions on the authenticity of audio and video files.

2.3. An Ensemble of Methods for Audio Tampering Detection

In the related work section, several approaches for tampering detection are presented, which may fall into specific categories. Such categories include relevant audible or inaudible artifacts that are produced during the malicious editing of audiovisual files. As a result, depending on the type of forgery and the technical flaws of such an action, one technique may be more or less suitable. Hence, the motivation of the project derives from the hypothesis that it is irrelevant to try to evaluate different approaches to choose the most efficient, since this cannot be applied universally to every case [8].

The MAthE AudioVisual Authentication approach proposes a superposition of methods in a modular architecture that includes a dynamic group of algorithmic elements. Such techniques are either outcomes of previous research work within the project [7,8,47] or were found in the literature. The modular architecture allows for the modification of existing functions in the future, as well as its extension with new modules that come from new research, literature review, or contribution within the academic society.

Another hypothesis that has played an important role in the MAthE architecture design is that the lack of real-world datasets, as well as the diversity of the characteristics of tampered files, sets a bottleneck to the maturity of automated decision-making schemes. Most models are trained with artificially created datasets that address a specific type of tampering (recording device, room acoustics, encoding, etc.). Moreover, disinformation is only relevant at certain time points of a file, where the editing alters the semantic meaning of the recording. This is something that a human subject may easily understand. For this reason, the proposed design incorporates signal processing tools and machine learning models in a semi-automated approach [48]. It is not within the project's expectations to provide automated massive authentication of archive files, but rather to assist humans in analyzing and authenticating a specific file under investigation (FUI). The outcomes of the system require a human-in-the-loop [49] strategy. This is considered an effective combination of machine processing capabilities and human intelligence.

The initial toolbox of signal processing algorithms that was included in the prototype version of the MAthE AudioVisual Authentication application is presented below. It is noted that the technical presentation and validation of every approach is not within the scope of the current paper. Instead, a short description of the main functional principles of every category of techniques is given, along with references to publications with the technical details of different algorithms. The toolbox is dynamic, and it will be supported by incorporating state-of-the-art feature-based [50,51] and deep [52] (machine) learning approaches for audiovisual semantic analysis. It can also be deployed as a mobile application [53]. It is expected to further grow and evolve through the use of the application and the continuous dissemination of the MAthE project.

2.3.1. Common Audio Representations

This family of tools includes typical audio representations, such as waveform, energy values, and spectrograms. Such tools are available in most typical audio editing applications. Sound waveforms are the depiction of the amplitude of sound pressure of every audio sample, expressing the variation in the audio signal in time. For an audio signal with a common sampling frequency of 44,100 samples per second, waveforms may include a huge number of samples to be depicted, which can be computationally heavy to represent in an interactive graph running on a web browser. For this reason, in the proposed toolbox, time integration is performed, showing the root mean square (RMS) value for successive time windows as a bar diagram. This is used as a tradeoff to avoid the excess information redundancy of the waveform. Mel scale spectrograms provide a spatiotemporal representation of audio signals, depicting the evolution of the spectral characteristics through a time interval [54]. Spectral information is given for specific frequency bands that apply to the Mel scale, which is inspired by the psychoacoustic characteristics of human auditory perception. They are included in the toolbox because they can be useful, and they enhance the MAthE framework's all-in-one solution so that users do not have to make use of more than one piece of software for analysis and decision-making.

2.3.2. Different Encoding Recognition

This family of techniques investigates the existence of small audio segments in the FUI that have different compression levels or encoding characteristics. This indicates that they may be segments of another file that were inserted in the original file. One common naïve approach that can be very effective in some cases is the calculation of the bandwidth, because most compression algorithms apply low-pass filtering to eliminate the higher frequencies that are of minor importance to the human auditory perception.

Feature vectors are descriptors of several attributes of a signal. Different encoding and compression levels, even if they are often proven to be inaudible in listening tests with human subjects, can affect the features that describe an audio signal. In the Double Compression technique for audio tampering detection that was proposed in [8], the FUI was heavily compressed. Features are extracted from the FUI and the compressed signal. For every time frame, the feature vector difference is calculated between the two signals. Parts of the FUI that have different encoding are expected to have different feature vector distances. Moreover, the gradient of differences is calculated. This measure is expected to reach peak values when there is an alteration in the compression levels, indicating suspicious points.

The double compression algorithm is summarized as follows [8]:

1. Heavy compression to the audio file under investigation (FUI), thus creating a double-compressed file (DCF).
2. A feature vector is extracted the FUI and the DCF, creating the ($T \times F$) matrices Fi (t), where i = 1, 2, T is the number of time frames and F is the length of the feature vector.
3. For every time frame, the Euclidean distance D(t) of the two matrices is calculated
4. $D'(t) = D(t) - D(t-1)$ is calculated to show the differentiation between successive time frames.
5. D′(t) is expected to present local extrema in time frames that include a transition between audio segments of different compression, indicating possible tampering points.

For the feature selection, an audio feature vector was evaluated in [7]. Using a dedicated dataset creation script, a set of audio files were created, containing audio segments of different compression formats and bitrates. Specifically, segments of mp3-compressed audio in different bitrates were inserted randomly within an uncompressed file containing speech. Subjective evaluation experiments with three experts in the field of media production indicated that human listeners failed completely to detect the inserted segments for mp3 bitrates above 96 kbps, while they recognized approximately 10% of the inserted segments for mp3s of 64 kbps [7]. The dataset that was created in [7], along with the script for customized dataset generation, are documented and provided pub-

licly at http://m3c.web.auth.gr/research/datasets/audio-tampering-dataset/ (accessed on 26 January 2022).

The selected feature set includes several frequency domain attributes, namely spectral brightness, with predefined threshold frequencies 500 Hz, 1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, 4000 Hz, and 8000 Hz, as well as rolloff frequencies, which are the upper boundary frequencies that contain energy ratios of 0.3, 0.5, 0.7, or 0.9 to the total signal energy, and spectral statistics (Spectral Centroid, Spread, Skewness, Kurtosis, Spectral Flatness, 13 Mel Frequency Cepstral Coefficients, Zero Crossing Rate, and RMS energy). The technical details of the aforementioned feature vectors are outside the scope of the current paper, but the methodology and feature evaluation process are presented thoroughly in [7].

2.3.3. Reverberation Level Estimation

Another indicator that several segments of a FUI may have been inserted from a different file is the effect of acoustic conditions on the recording. Every space has different reverberation levels that affect the recording. Especially since most newsworthy events are not recorded in ideal conditions of professional recording studios, a regression model based on a Convolutional Neural Network architecture [47] was trained using a big dataset of simulated reverberation to provide a numerical estimations of the Signal-to-Reverberation ratio for every audio segment. Segments with outlier values are possibly related to malicious audio splicing.

Convolutional Neural Networks (CNNs) are a type of deep learning architecture that have gained popularity in audio recognition and event detection tasks [55,56]. One main reason for their recent widespread is is that there is no need for a handcrafted feature vector. Instead, a visual representation of the audio information is fed to the networks as an image, and the input layers extract hierarchical features in an unsupervised manner during training. Different kinds of input have been evaluated for deep learning techniques, with spectrograms being the dominant approach [54].

Signal-to-Reverberation-Ratio (SRR) can be a useful attribute that can indicate audio slicing. SRR expresses the ratio of the energy of the direct acoustic field to the reverberation acoustic field. It is determined by the acoustic characteristics of the space of the recording, the positioning of the sound source and the recording device. The distance where the levels of the direct and the reverberation sound are equal (SRR = 1) is called the critical distance. At distances closer than the critical distance, we can assume SRR > 1, and at distances that are farther than the critical distance, SRR < 1. The critical distance itself depends on the room acoustic attributes.

For recordings that take place under different conditions, the SRR is expected to differ. When segments from different recordings are pieced together, it is possible to detect the inconsistency in their SRR, even if it is not audible by human listeners. Calculating the SRR for different time windows can provide another criterion for audio tampering detection.

In the proposed approach, a deep learning regression model is used for a data-driven estimation of the SRR, based on simulation data. A 3600-second-long audio file containing pink noise was created, using the Adobe Audition generator. Using the same software, reverberation was added to the file with different SRRs. Ten different SRRs were chosen, resulting in 11 audio files (including the original), producing a 39600-second-long dataset. The same source audio file was used for all SRRs, so that the model is trained to recognize the reverberation and not information related to the content of different audio streams. The selected SRRs are shown in Table 1.

Table 1. The different Signal-to-Reverberation Ratios that were used for the model training.

Signal (%)	100	90	80	70	60	50	40	30	20	10	0
Reverberation (%)	0	10	20	30	40	50	60	70	80	90	100

The dataset was used for the training of a CNN regression model. The output of the model is a continuous value from 0 (no reverberation) to 1 (only reverberation). The model architecture is provided in Table 2.

Table 2. The architecture and hyper-parameters of the CNN model for reverberation level estimation.

Layer	Type	Configuration
1	Convolutional 2D Layer	16 filters Kernel size = (3,3) Strides = (1,1)
2	Max Pooling 2D Layer	Pool size = (2,2)
3	Dropout	Rate = 0.25
4	Convolutional 2D Layer	32 filters Kernel size = (3,3) Strides = (1,1)
5	Max Pooling 2D Layer	Pool size = (2,2)
6	Dropout	Rate = 0.25
7	Convolutional 2D Layer	64 filters Kernel size = (3,3) Strides = (1,1)
8	Dropout	Rate = 0.25
9	Convolutional 2D Layer	128 filters Kernel size = (3,3) Strides = (1,1)
10	Convolutional 2D Layer	256 filters Kernel size = (3,3) Strides = (1,1)
11	Flatten Layer	
12	Dense Neural Network	Output weights = 64 Activation = ReLU L2 regularizer
13	Dense Neural Network	Output weights = 64 Activation = ReLU
14	Dropout	Rate = 0.25
15	Dense Neural Network	Output weights = 24 Activation = Linear

2.3.4. Silent Period Clustering

Besides room acoustics, the background noise also characterizes a recording. The environmental noise that is recorded in speech recordings is often inaudible, since it is mixed with a speech signal of a much higher level. However, in the small periods of silence that occur between words and syllables, the background noise signal is dominant. In the case of combining two or more recordings to create a tampered audio file, different background noise patterns may be distinguishable. Initial investigation has shown that by exporting a feature vector from small segments of silence (~25 ms) and providing them to a clustering algorithm, it is feasible to separate the different environmental audio classes that are present in an unsupervised way [8].

2.4. Crowdsourcing for Dataset Creation, Validation, and User Cooperation

Crowdsourcing is a methodology for distributed problem-solving that happens online by the collective intelligence of a community in a specific predefined direction set by an organization [49]. It has gained interest thanks to its efficiency and applicability in multiple domains and tasks [57–61]. In machine learning and automation, it has become very popular for collaborative problem solving and dataset formulation and validation [57,59]. Users are expected to participate according to intrinsic (fun, personal growth, etc.) and extrinsic (payment, rewards, etc.) motives [59,60].

Within the MAthE approach, crowdsourcing is promoted in several ways. First of all, crowdsourcing was promoted for the collaboration on the formulation of a database

of tampered audiovisual files. The importance of a real-world dataset for the training and evaluation of different machine learning and computer-assisted tampering detection approaches was highlighted in the previous sections. When a user provides a file for analysis in the web application, the file is stored temporarily to perform the analysis. After the results are provided, the user is asked for permission to save the file in our dataset. The file can be stored with or without accompanying metadata concerning the user's final decision. In case of a positive response, the file is stored.

Besides submitting files, users can also provide validation of existing files in the database. Such files may have been uploaded by other users but have no evaluation concerning their integrity, or may be files that are already annotated. It is common practice to include several annotators to strengthen the reliability and overall quality of the dataset.

In the case of crowdsourcing, the validation of files that have been uploaded by other users without an evaluation, collective intelligence, and collaboration are integrated into the framework. Users are encouraged to submit their files even if the analysis did not help them determine the authenticity of the file, to get help from other users. In case of a response, the uploader is informed about the other users' suggestions. This facilitates collaborative decision making, and is also a more efficient dataset creation strategy because files are not submitted only when the uploader can decide with confidence.

2.5. Common Use Case Scenarios

For a more efficient presentation of the functionality offered by the interface, the three most common use case scenarios are presented. In the results section, the implementation of the functionality in terms of the user experience (UX) choices, and the technical details are presented.

Scenario 1: A user submits a file and uses the toolbox to determine its authenticity.

In this common scenario, a user submits a file by uploading an audio or video file or by providing a YouTube link. After the analysis takes place on the server side, the visualizations are provided to the user. The user locates the points in time where inconsistencies are observed, listens to the audio, and makes a determination concerning the authenticity of the file, as was explained in Section 2.2 concerning the user-centered computer supported design. The user is then asked whether they are willing to contribute the file and their decision to the database. If they decide not to contribute, the file is deleted.

Scenario 2: A user is unable to decide and asks for help from the community.

The user submits the files, and, after they investigate the visualizations provided by the toolbox, they are unable to make a decision on the authenticity of the file. The user then decides to upload the file to the dataset unlabeled, so that it can be accessed by other users.

Scenario 3: A user browses the database to annotate files.

A user wants to contribute by annotating files that are already in the dataset. The interface provides randomly selected files from the database, along with the visualizations that come as outputs of the analysis. The user investigates the visualizations and makes a decision concerning the authenticity of the file, then chooses to submit their decision. Their decision is saved in the database, containing a label of the file (tampered/not tampered), and, optionally, the point in time where forging was detected and a short justification. The media file will still be available to other users for investigation after the annotation process. This means that a file may have multiple labels from different users, which is a common practice in crowdsourcing projects, since the input of one user cannot be considered totally reliable on its own.

3. Results

3.1. Convolutional Neural Network Regression Model for Signal-to-Reverberation-Ratio Estimation

As described in Section 2.2, among the integrated visualizations based on previous work, a CNN model was trained for the estimation of SRR from audio. The loss function

that was used for model training was mean square error (mse), which is a common choice for regression tasks, and Adamax was the optimizer. The goal of a regression training task is to minimize the difference between the estimated and the real values. The architecture and hyperparameters of the network were selected based on the existing literature and the trial-and-error-based micro-tuning during training. For the implementation, the Keras Python library was used [62]. Mel spectrograms were extracted to be used as input to the network, with 128 Mel scale coefficients. The spectrograms were extracted using the librosa library in Python [63]. They were extracted from overlapping windows of 1 s with a 50% overlap, leading to a final dataset of approximately 79,200 audio samples and a sampling frequency of 44,100 samples/second. For the output layer of the network, a fully connected network with a linear activation function was used to provide the predicted continuous value from 0 to 1. The mean square error was used as a metric for the evaluation of the network performance. A test set was used, that equals 20% of the entire dataset. The resulting mse was 0.029 (2.9%), a value that is considered acceptable for the task described.

3.2. Implementation and Deployment of the Prototype for Human-Centered Audio Authentication Support and Crowdsourcing

The web application was designed using the Flask web framework. This was a rational choice, taking into consideration the popularity of the Python programming language for data analysis and the successful deployment of similar web applications by our team [61]. The selection of a popular programming environment for such tasks may make the extendibility of the framework by contributors more appealing and viable. It was deployed on a dedicated Ubuntu virtual machine and was run on a Waitress production-quality pure-Python WSGI server. The application provides a back-end, where the algorithmic procedures take place, and a front-end graphical user interface.

For audio analysis, namely audio file read, write, and audio feature and spectrogram extractions, the librosa Python library was used. The PyTube Python library was used for YouTube video downloading. The AudioSegment Python library was used for mp3 transcoding, in order to implement the double compression algorithm. The CNN model for SRR estimation was saved as a TensorFlow HDF5 [64] model, and it was loaded during server launch in order to perform regression on the back end for the provided files.

The interface offers two main functionalities: Analyze and Contribute.

In Analyze mode, the user can provide an audiovisual object for investigation (Figure 2). The interface gives the choice of uploading a media file or providing a YouTube link. The analysis takes place on the server and returns a set of interactive visualizations based on the algorithmic procedures. As explained in Section 2.2, the framework follows a modular approach, allowing extension with more visualizations in future versions. The diagrams are generated using the Bokeh library. The main idea is the use of a linked x-axis for all figures, which is the time axis. This means that, by zooming in on a specific time value of one of the available diagrams, the user zooms automatically in on the same time value on all diagrams. This enables a simultaneous combined investigation of the results of all available algorithmic procedures for the detection of suspicious points within the file. For example, in Figure 3, by zooming in on a peak of the gradient of the double compression feature distances (upper right), which indicates a suspicious point, it is clear that the other diagrams also indicate a possible tampering point. The red circles noting the suspicious behavior in the three visualizations were added for presentation reasons in this paper and were not part of the original results by the interface. For further information concerning the principles of the aforementioned algorithmic procedures, refer to [8]. Moreover, a media player is provided, where the user can play the audio/video file at a certain point in time to assist with their decision-making.

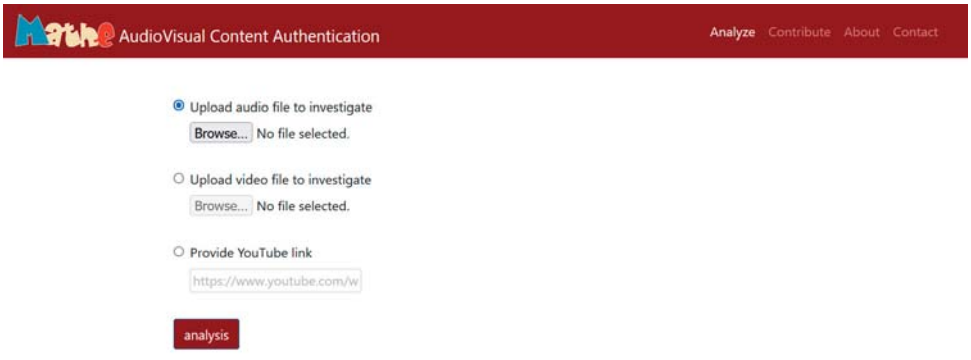


Figure 2. The interface where users can provide audio/video files for analysis.

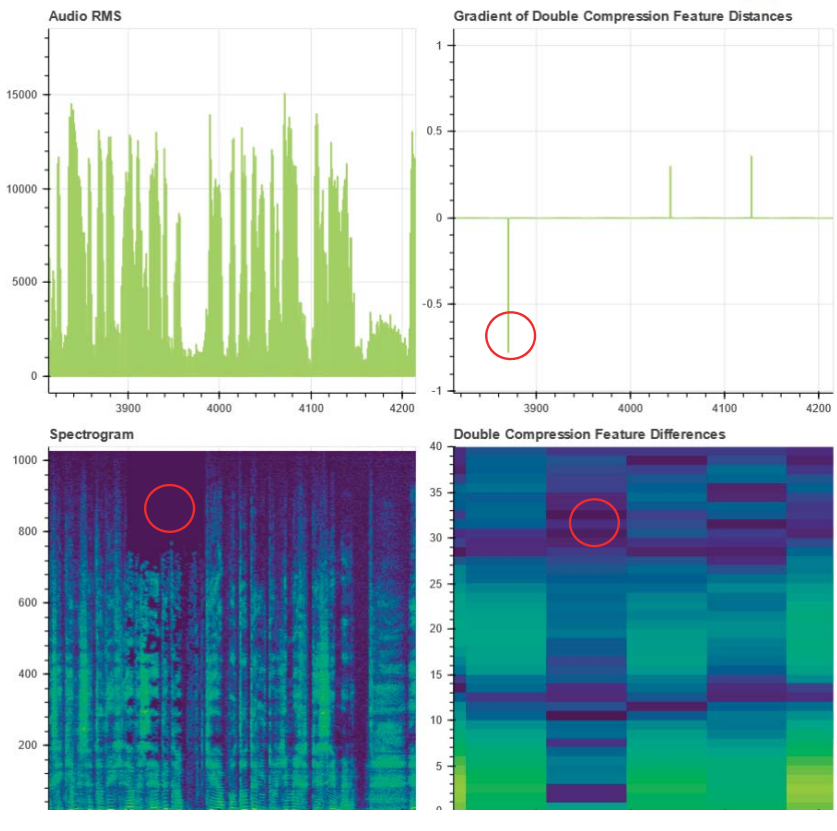


Figure 3. An example of combined analysis. The user has zoomed in at a suspicious point in time, and three of the visualizations indicate forgery.

The media file is uploaded in a temporary folder for the needs of analysis, to be deleted later. However, the application provides the user with the option to submit the file to the dataset, along with the analysis results and, optionally, a personal opinion on the authenticity of the file, as described in Section 2.3. The files, the analysis results, and the user annotations are stored in the database.

In the Contribute mode, users can annotate existing files of the database, in a crowdsourcing approach (Figure 4). A file is selected randomly. The same environment as in the Analyze mode is provided. The user has access to a media player, as well as the resulting set of interactive visualizations for the detection of possible tampering points. It is noted that the analysis results have already been saved to the database and are loaded directly. The analysis is a computationally complex procedure that requires time, which makes the annotation process much more time-consuming. After seeing the analysis, the user can determine whether and at which point in time the file is tampered, and can also optionally justify this decision with a short description. It is also possible to skip a certain file and select another one randomly. The user's opinion is stored in the database for the extension of the annotated dataset.

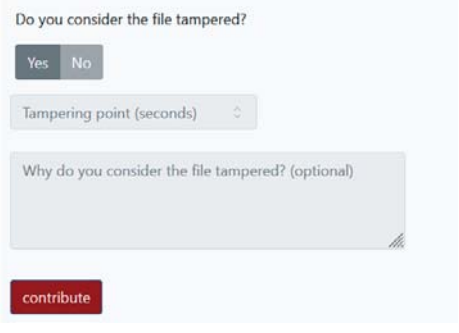


Figure 4. In Contribute mode, users can browse files from the database along with their analysis visualizations, and annotate them concerning the detection of audio tampering.

There is also an About section, for users who wish to get more information concerning the project and the specifications of the algorithmic implementations, and a Contact section for anyone who wishes to ask questions or contribute to the project. The current version of the web application is uploaded to the domain m3capps.jour.auth.gr (accessed on 26 January 2022) in testing mode, for evaluation.

It has been explained that contribution to the project is encouraged and sought after. Contribution cannot only be achieved through the use of the interface by users who want to submit files to the database or to annotate existing entries. It can also refer to providing new models or algorithms and improving the ones we have already incorporated, following the modular architecture that has been described in Section 3. To address such needs and also to strengthen the transparency of the proposed procedure, the code of the interface and the backend functionality has been uploaded to GitHub and can be retrieved at https://github.com/AuthJourM3C/MATHE-authentication (accessed on 26 January 2022) under a GNU General Public License v3.0.

4. Discussion

An AudioVisual Authentication application is presented, part of the MAthE project on computer-aided support of journalists and simple users against misinformation. It specializes in the authentication of audiovisual content in an audio-driven technical approach. It has the form of a web application and implements the functionality of a framework that promotes machine-assisted, human-centered decision making, collective intelligence, and collaboration in the battle against the malicious tampering of audiovisual content. The functionality of the application is provided to the end-users through a very simple and intuitive interface where the user is asked to provide the FUI. The toolbox features several signal processing modules that are applied to the FUI, providing an interactive graph that contains several visualizations. These are based on different technical principles and algorithms that aim to assist the user who makes the final decision. Through crowd-

sourcing, a dataset of real-world tampered files is expected to be created and validated for the first time.

Along with a set of algorithms that are based on previous research, a CNN regression model for SRR estimation was presented and evaluated. The model performs SRR estimation with an MSE of 0.029, which is an acceptable resolution for the detection of different acoustic environments. Of course, like all the proposed techniques, it has several limitations, especially regarding studio recordings, using files with similar room acoustics for tampering, or simulating the reverberation environment to match the excerpts used for copy-move forgery.

The main contributions of the research are summarized as follows:

1. A novel approach is proposed for audio tampering detection, where decision making is held by the human-in-the-loop in a computer assisted environment. This approach makes use of technical advances, surpasses their limitations and unreliability, and proposes a solution that can be immediately applied in journalistic practice.
2. The solution is provided openly as a service, allowing its use by journalists and the audience, without any limitations on their equipment or platform.
3. The application follows a modular approach. This means that the modules that are integrated in the prototype can be updated easily, and more modules can be added in the near future.
4. A CNN model for data-driven SRR estimation to be used in the direction of audio authentication was presented and evaluated.
5. A crowdsourcing approach was introduced for both user collaboration in media authentication and dataset creation and annotation. Users contribute with their effort to the extension of the dataset of tampered media files and also assist other users who request support in the authentication of specific files.

Since this is the initial launch of the application, future research goals include a thorough evaluation of the interface and the provided tools. This can be done in focus groups containing professionals and also publicly, to evaluate how a broader audience receives the application. Such workshops can also provide publicity for the project. Crowdsourcing is a core aspect of MAthE, so it is crucial to approach potential users and engage them in using the application to collect initial results. The major outcome is expected to be the dataset, which will be publicly available. After the formulation of the initial dataset, more intense experimentation on the applicability of different machine learning architectures can take place. Moreover, the involvement of more contributors from the engineering world (researchers, students, coders, etc.) can aid the improvement and extension of the provided toolbox. For this reason, all the necessary material will be also publicly accessible through the application's website.

5. Limitations

One major limitation concerning the current research is that the interface, at the time of publishing this paper, was an evaluation prototype, as was indicated in the title and clarified throughout the paper. As a result, it will be used for the dissemination and evaluation of the framework, so several things are expected to change, be added, or be modified in future versions. Moreover, as explained in Section 3, the implementation depends on several third-party components, such as pyTube for YouTube content downloading, and the YouTube API itself. Since such components can be modified without warning, the application will have to be maintained to follow the latest functionality, updates and syntax of every component that is used. From the prospective of UX design, an error page has been integrated into the application to handle such exceptions, and to provide contact information for troubleshooting and bug reporting. Since the applied procedures are computationally heavy and require significant resources to guarantee a fast response time, for the evaluation version there is a restriction on the allowed file size and YouTube video length. A restriction in the allowed file types has been also set. These restrictions are expected to be lifted when the application is in production.

Author Contributions: Conceptualization, N.V., A.K., R.K. and C.D.; methodology, N.V., L.V. and R.K.; software, N.V., L.V. and R.K.; validation, A.K., R.K. and C.D.; formal analysis, N.V. and R.K.; investigation, N.V., R.K., A.K. and C.D.; resources, N.V., R.K. and A.K.; writing—original draft preparation, N.V., A.K., L.V., R.K. and C.D.; visualization, N.V. and L.V.; supervision, R.K. and C.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Katsaounidou, A.N.; Dimoulas, C.A. Integrating Content Authentication Support in Media Services. In *Encyclopedia of Information Science and Technology*, 4th ed.; IGI Global: Hershey, PA, USA, 2018; pp. 2908–2919. [CrossRef]
2. Katsaounidou, A.; Dimoulas, C. The Role of media educator on the age of misinformation Crisis. In Proceedings of the EJTA Teachers' Conference on Crisis Reporting, Thessaloniki, Greece, 18–19 October 2018.
3. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2019. [CrossRef]
4. Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C.; Veglis, A. MAthE the game: A serious game for education and training in news verification. *Educ. Sci.* **2019**, *9*, 155. [CrossRef]
5. Katsaounidou, A.; Vryzas, N.; Kotsakis, R.; Dimoulas, C. Multimodal News authentication as a service: The "True News" Extension. *J. Educ. Innov. Commun.* **2019**, 11–26. [CrossRef]
6. Katsaounidou, A.; Gardikiotis, A.; Tsipas, N.; Dimoulas, C. News authentication and tampered images: Evaluating the photo-truth impact through image verification algorithms. *Heliyon* **2020**, *6*, e05808. [CrossRef] [PubMed]
7. Vryzas, N.; Katsaounidou, A.; Kotsakis, R.; Dimoulas, C.A.; Kalliris, G. Investigation of audio tampering in broadcast content. In Proceedings of the Audio Engineering Society Convention 144, Milan, Italy, 23–26 May 2018.
8. Vryzas, N.; Katsaounidou, A.; Kotsakis, R.; Dimoulas, C.A.; Kalliris, G. Audio-driven multimedia content authentication as a service. In Proceedings of the Audio Engineering Society Convention 146, Dublin, Ireland, 20–23 March 2019.
9. Bakker, P. New journalism 3.0—Aggregation, content farms, and Huffinization: The rise of low-pay and no-pay journalism. In Proceedings of the Future of Journalism Conference, Cardiff, UK, 8–9 September 2011.
10. Graves, L.; Cherubini, F. *The Rise of Fact-Checking Sites in Europe*; Reuters Institute for the Study of Journalism: Oxford, UK, 2016.
11. Bakir, V.; McStay, A. Fake news and the economy of emotions: Problems, causes, solutions. *Digit. Journal.* **2018**, *6*, 154–175. [CrossRef]
12. Verma, J.P.; Agrawal, S.; Patel, B.; Patel, A. Big data analytics: Challenges and applications for text, audio, video, and social media data". *Int. J. Soft Comput. Artif. Intell. Appl.* **2016**, *5*, 41–51. [CrossRef]
13. Vlachos, A.; Riedel, S. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, 26 June 2014; pp. 18–22.
14. Zampoglou, M.; Papadopoulos, S.; Kompatsiaris, Y. Large-scale evaluation of splicing localization algorithms for web images. *Multimed. Tools Appl.* **2017**, *76*, 4801–4834. [CrossRef]
15. Zampoglou, M.; Papadopoulos, S.; Kompatsiaris, Y. Detecting image splicing in the wild (web). In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops, Turin, Italy, 29 June–3 July 2015; pp. 1–6.
16. Sitara, K.; Mehtre, B.M. Digital video tampering detection: An overview of passive techniques. *Digit. Investig.* **2016**, *18*, 8–22. [CrossRef]
17. Grigoras, C.; Smith, J.M. Audio Enhancement and Authentication. In *Encyclopedia of Forensic Sciences*; Elsevier: Amsterdam, The Netherlands, 2013.
18. Maher, R.C. Audio forensic examination. *IEEE Signal Process. Mag.* **2009**, *26*, 84–94. [CrossRef]
19. Koenig, B.E. Authentication of forensic audio recordings. *J. Audio Eng. Soc.* **1990**, *38*, 3–33.
20. Zakariah, M.; Khan, M.K.; Malik, H. Digital multimedia audio forensics: Past, present and future. *Multimed. Tools Appl.* **2018**, *77*, 1009–1040. [CrossRef]
21. Gupta, S.; Cho, S.; Kuo, C.C.J. Current developments and future trends in audio authentication. *IEEE Multimed.* **2011**, *19*, 50–59. [CrossRef]
22. Rodríguez, D.P.N.; Apolinário, J.A.; Biscainho, L.W.P. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Trans. Inf. Forensics Secur.* **2010**, *5*, 534–543. [CrossRef]
23. Grigoras, C. Applications of ENF analysis in forensic authentication of digital audio and video recordings. *J. Audio Eng. Soc.* **2009**, *57*, 643–661.
24. Brixen, E.B. Techniques for the authentication of digital audio recordings. In Proceedings of the Audio Engineering Society Convention 122, Vienna, Austria, 5–8 May 2007.
25. Hua, G.; Zhang, Y.; Goh, J.; Thing, V.L. Audio authentication by exploring the absolute-error-map of ENF signals. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1003–1016. [CrossRef]

26. Malik, H.; Farid, H. Audio forensics from acoustic reverberation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 1710–1713.
27. Zhao, H.; Malik, H. Audio recording location identification using acoustic environment signature. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1746–1759. [CrossRef]
28. Buchholz, R.; Kraetzer, C.; Dittmann, J. Microphone classification using Fourier coefficients. In Proceedings of the International Workshop on Information Hiding, Darmstadt, Germany, 8–10 June 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 235–246.
29. Garcia-Romero, D.; Espy-Wilson, C.Y. Automatic acquisition device identification from speech recordings. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 1806–1809.
30. Hafeez, A.; Malik, H.; Mahmood, K. Performance of blind microphone recognition algorithms in the presence of anti-forensic attacks. In Proceedings of the 2017 AES International Conference on Audio Forensics, Arlington, VA, USA, 15–17 June 2017.
31. Malik, H. Acoustic environment identification and its applications to audio forensics. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1827–1837. [CrossRef]
32. Narkhede, M.; Patole, R. Acoustic scene identification for audio authentication. In *Soft Computing and Signal Processing*; Springer: Singapore, 2019; pp. 593–602.
33. Patole, R.K.; Rege, P.P.; Suryawanshi, P. Acoustic environment identification using blind de-reverberation. In Proceedings of the 2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, 19–21 December 2016; pp. 495–500.
34. Qiao, M.; Sung, A.H.; Liu, Q. MP3 audio steganalysis. *Inf. Sci.* **2013**, *231*, 123–134. [CrossRef]
35. Yang, R.; Shi, Y.Q.; Huang, J. Detecting double compression of audio signal. In *Media Forensics and Security II, Proceedings of the IS&T/SPIE Electronic Imaging, San Jose, CA, USA, 17–21 January 2010*; SPIE: Bellingham, WA, USA, 2010; Volume 7541, p. 75410K.
36. Liu, Q.; Sung, A.H.; Qiao, M. Detection of double MP3 compression. *Cogn. Comput.* **2010**, *2*, 291–296. [CrossRef]
37. Seichter, D.; Cuccovillo, L.; Aichroth, P. AAC encoding detection and bitrate estimation using a convolutional neural network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2069–2073.
38. Lacroix, J.; Prime, Y.; Remy, A.; Derrien, O. Lossless audio checker: A software for the detection of upscaling, upsampling, and transcoding in lossless musical tracks. In Proceedings of the Audio Engineering Society Convention 139, New York, NY, USA, 29 October–1 November 2015.
39. Gärtner, D.; Dittmar, C.; Aichroth, P.; Cuccovillo, L.; Mann, S.; Schuller, G. Efficient cross-codec framing grid analysis for audio tampering detection. In Proceedings of the Audio Engineering Society Convention 136, Berlin, Germany, 26–29 April 2014.
40. Hennequin, R.; Royo-Letelier, J.; Moussallam, M. Codec independent lossy audio compression detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 726–730.
41. Luo, D.; Yang, R.; Huang, J. Identification of AMR decompressed audio. *Digit. Signal Process.* **2015**, *37*, 85–91. [CrossRef]
42. Maung, A.P.M.; Tew, Y.; Wong, K. Authentication of mp4 file by perceptual hash and data hiding. *Malays. J. Comput. Sci.* **2019**, *32*, 304–314. [CrossRef]
43. Khan, M.K.; Zakariah, M.; Malik, H.; Choo, K.K.R. A novel audio forensic data-set for digital multimedia forensics. *Aust. J. Forensic Sci.* **2018**, *50*, 525–542. [CrossRef]
44. Gärtner, D.; Cuccovillo, L.; Mann, S.; Aichroth, P. A multi-codec audio dataset for codec analysis and tampering detection. In Proceedings of the Audio Engineering Society Conference: 54th International Conference: Audio Forensics: Techniques, Technologies and Practice, London, UK, 12–14 June 2014.
45. Imran, M.; Ali, Z.; Bakhsh, S.T.; Akram, S. Blind detection of copy-move forgery in digital audio forensics. *IEEE Access* **2017**, *5*, 12843–12855. [CrossRef]
46. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
47. Vryzas, N. Audiovisual Stream Analysis and Management Automation in Digital Media and Mediated Communication. Ph.D. Dissertation, Aristotle University of Thessaloniki, Thessaloniki, Greece, 2020.
48. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D Deep CNNs vs. Temporal Feature Integration for General Audio Classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [CrossRef]
49. Brabham, D.C. *Crowdsourcing*; MIT Press: Cambridge, MA, USA, 2013.
50. Vrysis, L.; Hadjileontiadis, L.; Thoidis, I.; Dimoulas, C.; Papanikolaou, G. Enhanced Temporal Feature Integration in Audio Semantics via Alpha-Stable Modeling. *J. Audio Eng. Soc.* **2021**, *69*, 227–237. [CrossRef]
51. Bountourakis, V.; Vrysis, L.; Konstantoudakis, K.; Vryzas, N. An enhanced temporal feature integration method for environmental sound recognition. *Acoustics* **2019**, *1*, 410–422. [CrossRef]
52. Vrysis, L.; Thoidis, I.; Dimoulas, C.; Papanikolaou, G. Experimenting with 1D CNN Architectures for Generic Audio Classification. In Proceedings of the Audio Engineering Society Convention 148, Vienna, Austria, 2–5 June 2020.
53. Vrysis, L.; Vryzas, N.; Sidiropoulos, E.; Avraam, E.; Dimoulas, C.A. jReporter: A Smart Voice-Recording Mobile Application. In Proceedings of the Audio Engineering Society Convention 146, Dublin, Ireland, 20–23 March 2019.

54. Korvel, G.; Treigys, P.; Tamulevicus, G.; Bernataviciene, J.; Kostek, B. Analysis of 2d feature spaces for deep learning-based speech recognition. *J. Audio Eng. Soc.* **2018**, *66*, 1072–1081. [CrossRef]
55. Ciaburro, G. Sound event detection in underground parking garage using convolutional neural network. *Big Data Cogn. Comput.* **2020**, *4*, 20. [CrossRef]
56. Ciaburro, G.; Iannace, G. Improving smart cities safety using sound events detection based on deep neural network algorithms. *Informatics* **2020**, *7*, 23. [CrossRef]
57. Estellés-Arolas, E.; González-Ladrón-de-Guevara, F. Towards an integrated crowdsourcing definition. *J. Inf. Sci.* **2012**, *38*, 189–200. [CrossRef]
58. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Crowdsourcing audio semantics by means of hybrid bimodal segmentation with hierarchical classification. *J. Audio Eng. Soc.* **2016**, *64*, 1042–1054. [CrossRef]
59. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Mobile audio intelligence: From real time segmentation to crowd sourced semantics. In Proceedings of the Audio Mostly 2015 on Interaction with Sound, Thessaloniki, Greece, 7–9 October 2015; pp. 1–6.
60. Cartwright, M.; Dove, G.; Méndez Méndez, A.E.; Bello, J.P.; Nov, O. Crowdsourcing multi-label audio annotation tasks with citizen scientists. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–11.
61. Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. [CrossRef]
62. Chollet, F.; Eldeeb, A.; Bursztein, E.; Jin, H.; Watson, M.; Zhu, Q.S. *Keras*; (v.2.4.3); GitHub: San Francisco, CA, USA, 2015; Available online: https://github.com/fchollet/keras (accessed on 26 January 2022).
63. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in Python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
64. Collette, A. *Python and HDF5: Unlocking Scientific Data*; O'Reilly Media, Inc.: Newton, MA, USA, 2013.

future internet

MDPI

Article

MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation

Olga Papadopoulou [1,*,†], Themistoklis Makedas [1,†], Lazaros Apostolidis [1], Francesco Poldi [2], Symeon Papadopoulos [1] and Ioannis Kompatsiaris [1]

[1] Centre for Research and Technology Hellas—CERTH, Information Technologies Institute—ITI, 6th km Harilaou-Thermi, Thermi, 57001 Thessaloniki, Greece; tmakedas@iti.gr (T.M.); laaposto@iti.gr (L.A.); papadop@iti.gr (S.P.); ikom@iti.gr (I.K.)
[2] EU DisinfoLab, Chaussée de Charleroi 79, 1060 Brussels, Belgium; fp@disinfo.eu
* Correspondence: olgapapa@iti.gr; Tel.: +30-2311-257766
† These authors contributed equally to this work.

Abstract: The proliferation of online news, especially during the "infodemic" that emerged along with the COVID-19 pandemic, has rapidly increased the risk of and, more importantly, the volume of online misinformation. Online Social Networks (OSNs), such as Facebook, Twitter, and YouTube, serve as fertile ground for disseminating misinformation, making the need for tools for analyzing the social web and gaining insights into communities that drive misinformation online vital. We introduce the MeVer NetworkX analysis and visualization tool, which helps users delve into social media conversations, helps users gain insights about how information propagates, and provides intuition about communities formed via interactions. The contributions of our tool lie in easy navigation through a multitude of features that provide helpful insights about the account behaviors and information propagation, provide the support of Twitter, Facebook, and Telegram graphs, and provide the modularity to integrate more platforms. The tool also provides features that highlight suspicious accounts in a graph that a user should investigate further. We collected four Twitter datasets related to COVID-19 disinformation to present the tool's functionalities and evaluate its effectiveness.

Keywords: social network analysis; network visualization tools; online disinformation; online social networks; journalistic practices; intelligent metadata processing

Citation: Papadopoulou, O.; Makedas, T.; Apostolidis, L.; Poldi, F.; Papadopoulos, S.; Kompatsiaris, I. MeVer NetworkX: Network Analysis and Visualization for Tracing Disinformation. *Future Internet* **2022**, *14*, 147. https://doi.org/10.3390/fi14050147

Academic Editor: Eirini Eleni Tsiropoulou

Received: 9 April 2022
Accepted: 3 May 2022
Published: 10 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The increasing digitalization of our world offers significant opportunities for groundbreaking investigative journalism, new models of cross-border collaborative reporting, and access to treasure troves of knowledge and diverse sources at a mouse-click [1]. However, journalists struggle every day to cope with the overwhelming amount of information that emerges online. This combined with a time pressure to verify information as quickly as possible has caused a need for tools that can provide automatic or semi-automatic assistance to arise. Social Network Analysis (SNA) is a field that researchers turn to in order to build tools that can assist journalists in investigating topics disseminated through social media platforms by observing the propagation of claims and rumors, the discussions around the claims and rumors, and interactions between users. Coupled with intelligent methods that extract and process metadata, these tools can provide disinformation-related cues to journalists and fact-checkers and become vital in the daily activities of these professionals.

Networks are complex systems of actors, referred to as nodes, interconnected via relationships called edges. On social media, a node can be an account (i.e., a user, page, or group), a URL (i.e., an article or media item), or a keyword (i.e., a hashtag). When two nodes interact (i.e., when a Twitter account retweets a tweet of another Twitter account), an edge is formed between them. The usefulness of network visualizations is to investigate trends

and events as a whole. The challenging part of analyzing a social network is identifying the nodes and relationships that are worth investigating further.

An essential feature that makes social network analysis important for combating disinformation is that false news spreads faster than real news through online platforms involving many users and creating large networks [2]. For example, a CBC journalist [3] posted a wrong claim that identified the attacker of an attack in Toronto in 2018 as "angry" and "Middle Eastern" at the same time as another journalist who posted a claim correctly identifying the attacker as "white". It turned out that the misleading tweet identifying the attacker as Middle Eastern received far more engagement than the accurate one roughly five hours after the attack. A network emerged rapidly around the false claim, and users were quick to disseminate it. The visualization of a network involving many accounts and their interactions may reveal those accounts that try to influence the public with certain views.

During critical events, such as the 2016 US presidential election and the outbreak of the COVID-19 pandemic, fabricated information was disseminated through social media to deceive the public. Several works revealed the roles of bots (i.e., automated accounts posing as real users) in the spread of misinformation [2,4]. Their characteristics were excessive posting via the retweeting of emerging news and tagging or mentioning influential accounts in the hope they would spread the content to their thousands of followers [5]. A need to detect inauthentic users led to investigating the posting activities, interactions, and spreading behaviors. Network analysis and visualization techniques could be valuable for detecting such inauthentic accounts based on their behaviors in a network that differentiate them from those of real users.

In this work, we present the MeVer NetworkX analysis and visualization tool. The tool's development was motivated by a need to follow complex social media conversations and to gain insights about how information is spreading in networks and how groups frequently communicate with each other and form communities. The tool falls in the scope of assisting journalistic practices and, more precisely, helping journalists retrieve specific and detailed information or form a comprehensive view around a complex online event or topic of discussion. The tool aggregates publicly available information on accounts and disseminated messages and presents them in a convenient semantically enriched network view that is easy to navigate and filter, aiming to overcome the critical challenge of unstructured data on the Web. We focused on implementing a clear and straightforward navigation with no overlap among communities to provide users with easy-to-digest visualizations. A multitude of implemented features provide insights into the propagation flows and the behaviors of accounts. The primary functionality of the tool is to highlight suspicious accounts worth investigation, which could potentially speed up manual analysis processes. Finally, the tool is among the few to support three social media platforms (Twitter, Facebook, and Telegram), and its modular nature makes it extensible to more platforms. With this work, we aim to leverage intelligent metadata extractions, processing, and network science to endow journalists and fact-checkers with advanced tools in their fight against disinformation.

2. Related Work

A significant challenge in creating useful social graphs for the analysis of online phenomena relates to the process of data collection. The challenge is that users need to have specialized knowledge to collect data and that platforms have limitations on available data. Another aspect that strictly relates to social graph analysis is the field of bot/spammer detection. A common tactic for spreading disinformation quickly and widely is to create fake accounts that pretend to be authentic. Research in this field revealed that these accounts have certain characteristics and behaviors that lead to their automatic detection. In the following sections, we list the visualization tools introduced in the literature with a brief description of their functionalities and limitations.

2.1. Data Collection and Analysis

A prerequisite for creating social graphs is the collection of actors (nodes) and relationships (edges) for a query topic. In online social networks, such as Facebook and Twitter, actors can be accounts, links, hashtags, and others and edges can represent connections.

The Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT) [6] is a toolset for capturing and analyzing Twitter data. It relies on the Twitter search API to download tweets (from the last 7 days due to Twitter's rate limits) based on a search term. It provides some basic statistics on a collected dataset (the number of tweets with URLs, hashtags, and mentions; the number of tweets/retweets; and the numbers of unique users in the dataset). It creates different networks (users, co-hashtags, users–hashtags, and hashtags–URLs) and supports exporting them to the GEXF (Graph-Exchange XML Format) for visualizations. Similarly, for Twitter data, a component called Twitter SNA [7] was developed as part of the InVID-WeVerify verification plugin [8], which supports the collection of Twitter data. A plugin component transforms collected data into a format that is suitable for network visualizations and supports exports to the GEXF. CrowdTangle (https://www.crowdtangle.com/ accessed on 8 April 2022) is a tool that supports data collection for building Facebook graphs. It provides a user with an export functionality with which posts by public Facebook pages and groups are listed and accompanied by metadata. While most tools are built focusing on the collection of data from a specific platform, the open-source 4CAT Capture and Analysis Toolkit [9] (4CAT) can capture data from a variety of online sources, including Twitter, Telegram, Reddit, 4chan, 8kun, BitChute, Douban, and Parler.

An advantage of the presented tool is that it is already integrated with the Twitter SNA component (which is currently accessible through authentication and is reserved for fact checkers, journalists, and researchers to avoid misuse) of the InVID-WeVerify verification plugin so that users can automatically trigger 4CAT with query campaigns they want to investigate.

2.2. Bot/Spammer Detection

Significant research has been conducted to identify the spread of disinformation and spam on OSNs, especially on Twitter. Recent work proposed features based on an account's profile information and posting behaviors and applied machine-learning techniques to detect suspicious accounts. The authors in [10] examined tweet content itself and included information about an account that posted a tweet as well as n grams and sentiment features in order to detect tweets carrying disinformation. Similarly, in [11], the authors attempted to find the minimum best set of features to detect all types of spammers. In [12], a hybrid technique was proposed that uses content- and graph-based features for the identification of spammers on the platform Twitter. In [13], the authors proposed various account-, content-, graph-, time-, and automation-based features, and they assessed the robustness of these features. Other similar machine-learning techniques were proposed in [14–16]. In [17], the authors focused on the detection of not just spam accounts but also on regular accounts that spread disinformation in a coordinated way. In [18], a different methodology was followed based on a bipartite user–content graph. This work assumed that complicit spammers need to share the same content for better coverage. Shared content is also a more significant complicity signal than an unsolicited link on Twitter. The user similarity graph consisted of nodes as users and edges that represented the similarity between the users. Finally, a complete survey of recent developments in Twitter spam detection was presented in [19]. A proposed tool provided users with a convenient mechanism for inspecting suspicious accounts, leveraging features introduced in the literature. However, the automatic algorithms for detecting spam accounts are not yet part of the tool.

2.3. Visualization Tools

One of the most popular and most used open-source software options for network visualization and analysis is Gephi (https://gephi.org/ accessed on 8 April 2022). It

provides a multitude of functionalities for the easy creation of social data connectors to map community organizations and small-world networks. Gephi consists of many functionalities that provide users the ability to visualize very large networks (up to 100,000 nodes and 1,000,000 edges) and to manipulate the networks using dynamic filtering and SNA methods. Other network visualization tools include GraphVis (https://networkrepository.com/graphvis.php accessed on 8 April 2022), which is for interactive visual graph mining and relational learning, and webweb, (https://webwebpage.github.io/ accessed on 8 April 2022) which is for creating, displaying, and sharing interactive network visualizations on the web. ORA is a toolkit for dynamic network analyses and visualizations that supports highly dimensional network data. It is a multi-platform network toolkit that supports multiple types of analyses (e.g., social network analyses using standard social network metrics; examinations of geo-temporal networks; identifications of key actors, key topics, and hot spots of activity; and identifications of communities). NodeXL is an extensible toolkit for network overviews, discoveries, and exploration and is implemented as an add-on to the spreadsheet software Microsoft Excel 2007. It supports both the data import process and analysis functionalities, such as the computation of network statistics and the refinement of network visualizations through sorting, filtering, and clustering functions [20].

A recently introduced open-source interface for scientists to explore Twitter data through interactive network visualizations is the Twitter Explorer [21]. It makes use of the Twitter search API with all the limitations (number of requests per 15 min and tweets from the last seven days) to collect tweets based on a search term and analyze them. It includes a Twitter timeline of the collected tweets, creates interaction networks and hashtag co-occurrence networks, and provides further visualization options. A tool that visualizes the spread of information on Twitter is Hoaxy (https://hoaxy.osome.iu.edu/ accessed on 8 April 2022). This lets users track online articles posted on Twitter, but only those posted within the last seven days. A user can add a search term and visualize the interactions of at most 1000 accounts that share the term. This tool creates a graph in which each node is a Twitter account and two nodes are connected if a link to a story passes between those two accounts via retweets, replies, quotes, or mentions. Hoaxy uses the Botometer score for coloring the nodes, which calculates the level of automation an account presents using a machine-learning algorithm trained to classify.

Looking into more specialized tools, Karmakharm et al. [22] presented a tool for rumor detection that can continuously learn from journalists' feedback on given social media posts through a web-based interface. The feedback allows the system to improve an underlying state-of-the-art neural-network-based rumor classification model. The Social Media Analysis Toolkit [23] (SMAT) emerged from the challenge of dealing with the volume and complexity of analyzing social media across multiple platforms, especially for researchers without computer science backgrounds. It provides a back end data store that supports different aggregations and supports exporting results to easy user-friendly interfaces for fast large-scale exploratory analyses that can be deployed on a cloud. Significant research has focused on misinformation on the Twitter platform. BotSlayer [24] is a tool that detects and tracks the potential amplification of information by bots on Twitter that are likely coordinated in real time. Reuters Tracer [25] is a system that helps sift through noise to detect news events and assess their veracities. Birdspotter [26] aims to assist non-data science experts in analyzing and labeling Twitter users by presenting an exploratory visualizer based on a variety of computed metrics.

Our proposed tool provides several functionalities that are similar or complementary to the existing tools. Later in the paper, we present a comparison of our tool with Gephi and Hoaxy.

3. MeVer NetworkX Tool

The proposed tool is a web-based application for the visualization of Twitter, Facebook, and Telegram graphs. Each user submits an input file and provides their email, and a link

to a resulting graph is sent to them as soon as the processing of that file has been completed. This asynchronous means of returning the results is considered more acceptable by users, especially in cases of large graphs for which the processing time is long (several minutes). Figure 1 illustrates an overview of the tool, which is mainly separated into a server-side analysis and user interface.

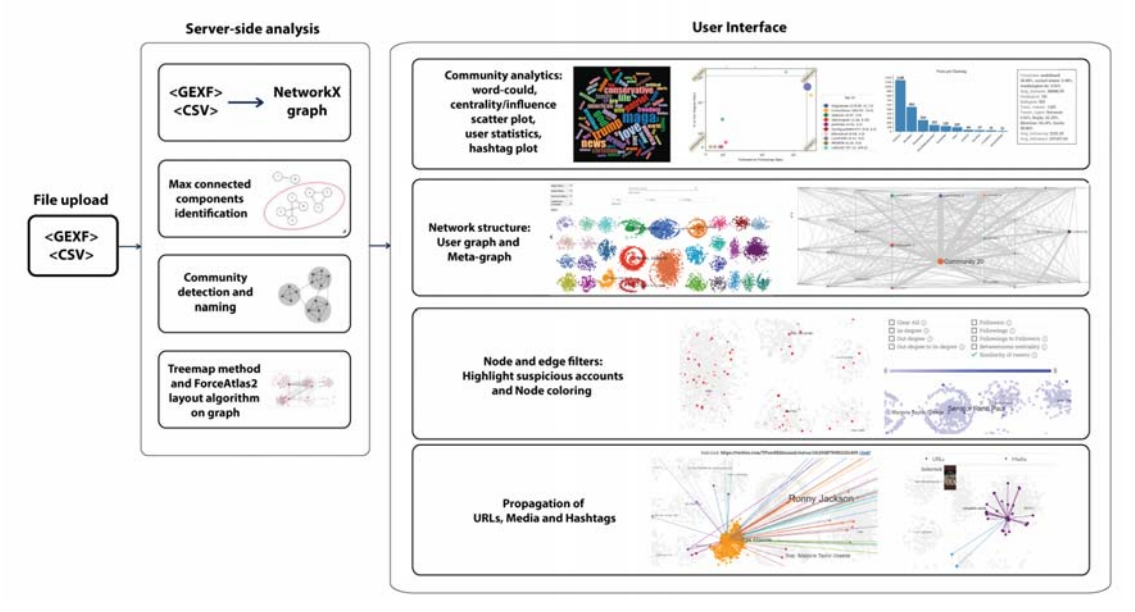


Figure 1. Overview of the MeVer NetworkX tool.

3.1. Input Files

Twitter and Telegram analyses involve GEXF files. This well-known format started in 2007 within the Gephi project to describe complex network structures and their associated data and dynamics. To build Facebook graphs, a user needs access to CrowdTangle and then he/she can export the data in CSV files.

Twitter. The input GEXF file of a Twitter graph contains the nodes and edges of the graph along with their attributes. The required field for a node is an ID, and the edges need the source of the edge (i.e., the ID of the node from which the edge starts and the "target" of the edge (i.e., the node's id to which the edge points)). Additional attributes are used to build plots, statistics, filters, and other functionalities. The graph contains three types of nodes visualized with different shapes: users visualized as circles, URLs visualized as stars, and hashtags visualized as rhombuses. The tool supports four edge types: retweets, quotes, mentions, and replies. Table 1 presents a list of required and optional fields. Account-related attributes (e.g., screen names and numbers of accounts following) are associated with nodes, while tweet-related attributes (e.g., retweets, texts, and hashtags) are associated with edges and characterize the interactions between nodes.

Table 1. Twitter's attributes for building the graph and features. All fields marked with an asterisk (*) are required.

Attribute	Description	Type
Node ID *	Unique ID	Node
Type of node	Set to true if the node is a user node	Node
Tweet ID	Unique tweet ID	Node
Screen name	Screen name, handle, or alias that a user identifies themselves as; screen_names are unique but subject to change.	Node
Created at	UTC time when this tweet was created. Example:	Node
User description	User-defined UTF-8 string describing their account.	Node
Names	Name of the user as they have defined it in Twitter.	Node
Number of followers	Number of followers this account currently has	Node
Location	User-defined location for this account's profile.	Node
Number of accounts following	Number of users this account is following	Node
Verified	When true, indicates that the user has a verified account.	Node
Number of statuses	Number of tweets (including retweets) issued by the user.	Node
Profile image	HTTPS-based URL pointing to the user's profile image.	Node
Background image	HTTPS-based URL pointing to the standard Web representation of the user's uploaded profile banner.	Node
Edge ID	Unique ID	Edge
Source *	Node that the edge starts at	Edge
Target *	Node that the edge points to	Edge
Tweet ID	Unique tweet ID	Edge
Retweet	Whether the edge is a retweet	Edge
Reply	Whether the edge is a reply	Edge
Mention	Whether the edge is a mention	Edge
Quote	Whether the edge is a quote	Edge
Created at	UTC time when this tweet was created.	Edge
Number of retweets	Number of times this tweet has been retweeted.	Edge
Number of favorites	Approximately how many times this tweet has been liked by Twitter users.	Edge
Text	Actual UTF-8 text of the status update.	Edge
Hashtags	Hashtags that have been parsed out of the tweet text.	Edge
URLs	URLs included in the text of a tweet.	Edge
Media	Media elements uploaded with the tweet.	Edge

Facebook. CrowdTangle tracks only publicly available posts and extracts data in CSV files. The CSV files contain one public Facebook post per line with metadata about the page/group that posted it and the post itself. We considered two types of node: groups and resources (URLs, photos, or videos); the interactions among them are defined as the edges of a graph. The nodes are visualized with different shapes, namely a circle for a Facebook group/page, a star for an article, a rhombus for a photo, and a square for a video. An edge is created from a group/page node made into a resource node when the group/page shares a post containing the resource node's link. When multiple groups share resources, multiple edges are created for the resource node. Metadata that refers to Facebook pages/groups are used as node attributes, while information related to Facebook posts that contain resources (URLs, photos, or videos) is associated with edges since a resource might be associated with multiple pages/groups. Table 2 summarizes and explains the attributes used for building Facebook graphs.

Telegram. The Telegram graph has three node types visualized, each with a different shape: (i) users (circles), (ii) URLs (stars), and (iii) hashtags (rhombuses). Edges represent occurrences of hashtags and URLs within the text field of a message sent by a user through a channel. Based on how the Telegram ecosystem is conceived and constructed, it's not possible to determine the actual Telegram account that used the channel as a mean to communicate with its subscribers. However, channel administrators can use the author's signature feature in order to include the first and last name in the signature of each message they send. Such an indicator cannot lead anyone to a unique identification of the Telegram account that has been used to send certain messages containing a specific signature. Due to

this, the user node visualized with a circle in the graph corresponds to the channel name and not the author of the messages.

Table 2. Facebook's attributes for building the Facebook graph and features. All fields marked with an asterisk (*) are required.

Attribute	Description	Type
Page/group name	Name of the page/group that posted	Node
User name	Username of the page/group	Node
Facebook ID *	ID of the page/group	Node
Likes at posting	Number of likes of the page/group at the time of posting	Node
Followers at posting	Number of page/group followers at the time of posting	Node
Type	Types of links (articles, photos, and videos) included in the post	Node
Resource *	Link included in the post	Node/edge
Total interactions	Total number of all reactions (likes, shares, etc.)	Edge
Message	Message written in the post	Edge
Created	Time the post was published	Edge
Likes	Number of likes on the post	Edge
Comments	Number of comments on the post	Edge
Shares	Number of shares of the post	Edge
Love	Number of love reactions on the post	Edge
Wow	Number of wow reactions on the post	Edge
Haha	Number of haha reactions on the post	Edge
Sad	Number of sad reactions on the post	Edge
Angry	Number of angry reactions on the post	Edge
Care	Number of care reactions on the post	Edge

At this point, it is really important to underline and remark that the researchers anonymized the data before operating or storing it with a cryptographic hash function named *BLAKE2*, which is defined in RFC 7693 (https://datatracker.ietf.org/doc/html/rfc7693.html accessed on 8 April 2022). If a message contains one or more hashtags and/or one or more URLs, a node is created for each entity that can be extracted, and an edge connecting each couple of nodes is created too. As per our ethical considerations, only open-source and publicly available information was gathered and analyzed. This ethical deliberation should not be interpreted as an actual obstacle or limit, given that disinformation actors prefer these public Telegram venues. The entities that can be extracted from each message are listed in Table 3.

Table 3. Telegram's attributes for building the Telegram graph and features. The field marked with an asterisk (*) is optional.

Attribute	Description	Type
ID	Unique identifier of the message within the channel	Edge
Message link	URL to the message in the object	Edge
Hashtags	Hashtags included in the message	Node/edge
Links	Links included in the message	Node/edge
Timestamp	The time at which the message was sent	Edge
Message	Text of message	Edge
Author's signature *	First and last name of the author of the message	Edge
Views	Number of times a message was viewed	Edge

3.2. Features and Functionalities

3.2.1. Individual Account and Post Inspections

Users can click on individual accounts of interest and obtain information about various account statistics, the most influential nodes they are connected to, and the communities they interact with the most (Figure 2a). Additionally, a user can focus the visualization on

the interactions (edges) of an account in a graph (Figure 2c). Finally, the text of the posts (only for Twitter) made by a selected user are presented (Figure 2b).

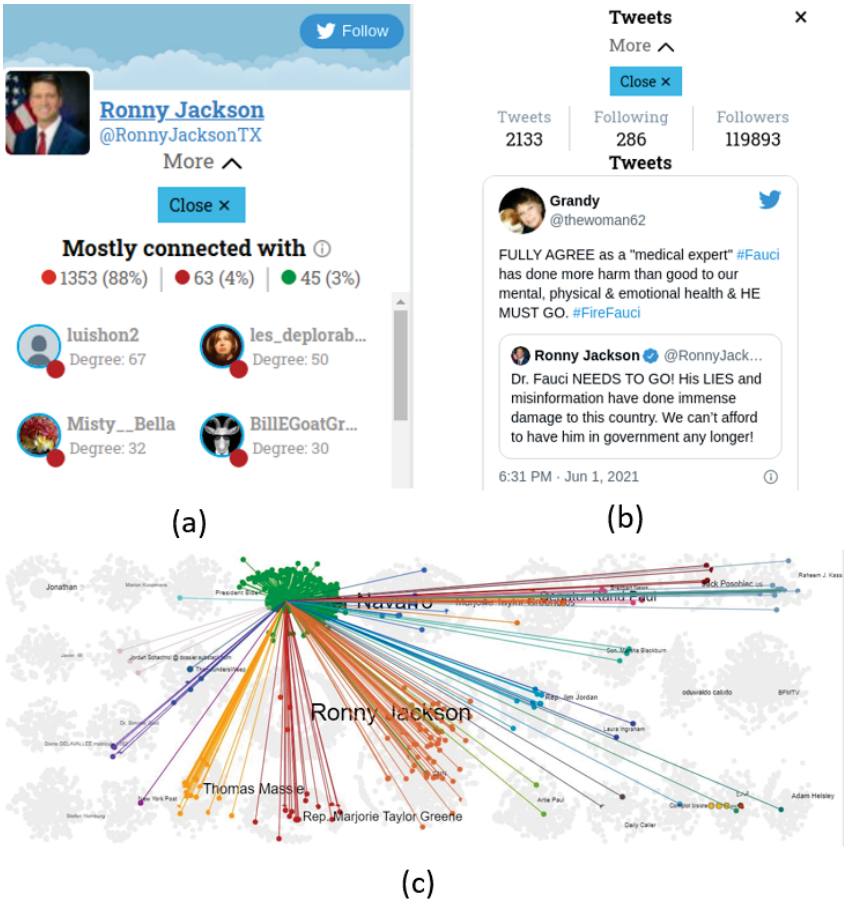


Figure 2. Individual user inspection. (**a**) Information about the user, (**b**) text of the tweets made by the user and (**c**) the interactions of this user.

3.2.2. Community Detection

We identified all the connected components and then discarded those with fewer than three nodes. We undertook this mainly because in the vast majority of cases, communities with very few nodes do not provide helpful insights on a network. In such situations, the Louvain [27] algorithm was applied to perform community detection on a graph. The formed communities were named using the names of the top three nodes based on their degrees (number of edges a node has). Nodes with high degrees can provide insights about the rest of a community as they are usually centered around a specific topic. The last step before the visualization refers to the positioning of nodes within communities and then positioning the communities in the graph. To this end, we followed a two-step procedure that is an adaptation of the TreeMap methodology [28] in combination with the ForceAtlas2 [29] layout algorithm, which prevents overlapping nodes and provides a clear and readable graph. The main idea of the TreeMap positioning method was the division of the screen area into rectangles of different sizes and the assignment of each

community into a rectangle taking into account the number of nodes that belonged to the corresponding community. To adapt this method to our needs, we implemented Algorithm 1 to order communities in order to position the largest community at the center of the screen and the rest of the communities around it, taking into account the number of interactions and avoiding overlapping. Next, Algorithm 2 allocated the screen area to each community so that larger communities spanned larger areas. Algorithm 3 performed the above procedures and executed a Python implementation of the *TreeMap* algorithm to position the communities and then executed a Python implementation of the ForceAtlas2 (https://github.com/bhargavchippada/forceatlas2 accessed on 8 April 2022) layout algorithm to position the nodes within the communities.

Algorithm 1 Order communities. Largest community at the center of the graph. Communities with most interactions are closest.

```
procedure COMMUNITY_LIST(coms, edges)
    k ← 0
    max_com_id ← get_largest_com(coms)
    middle ← len(coms)/2
    order_coms[middle] ← coms[max_com_id]
    temp_id ← max_com_id
    while len(coms) > 0 do
        for i in edges do
            max_edges_id ← find_max_edges(i, temp_id)
        end for
        k ← k + 1
        if add to left then
            order_coms[middle − k] ← coms[max_edges_id]
        else
            order_coms[middle + k] ← coms[max_edges_id]
        end if
        temp_id ← max_edges_id
        remove(coms[max_edges_id])
    end whilereturn order_coms
end procedure
```

Algorithm 2 Allocation of areas to communities based on the communities' sizes.

```
procedure ALLOCATE_AREA(order_coms_sizes, width, height)
    minimum_size ← round(width ∗ height/len(order_coms_sizes)/2)
    total_area ← width ∗ height − minimum_size ∗ len(order_coms_sizes)
    for size in order_coms_sizes do
        com_sizes ← (size ∗ total_area/total_size) + minimum_size
    end forreturn com_sizes
end procedure
```

Algorithm 3 Positioning of communities and nodes.

```
procedure POSITIONING(coms, edge)
    width ← Screen_width
    height ← Screen_height
    order_coms ← community_list(coms, edge)
    com_sizes ← allocate_area(order_coms_sizes, width, height)
    rectangles ← treemap(order_coms, com_sizes)
    x_node, y_node ← ForceAtlas2(rectangles) return x_node, y_node
end procedure
```

Figure 3 shows an example graph. A user can obtain insights about the number of nodes and edges in a graph and easily identify the most influential nodes since they are the largest ones and their names are shown.

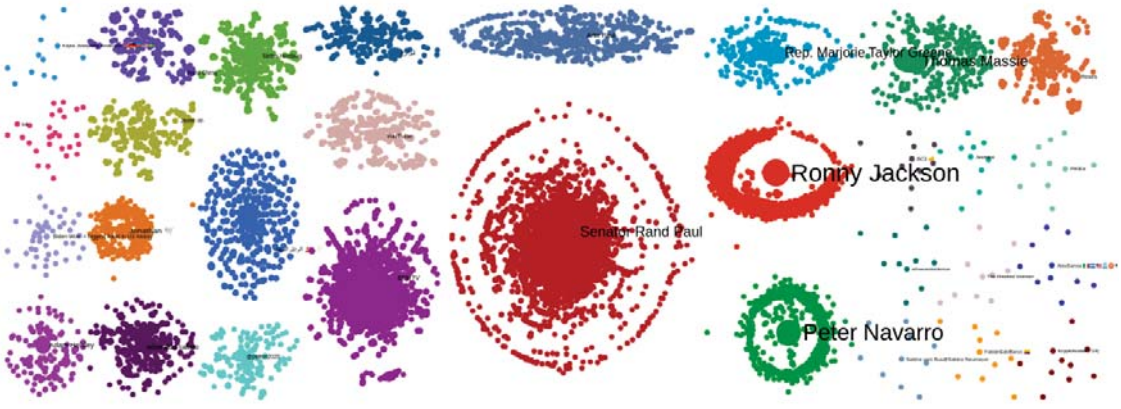


Figure 3. Visualization of example graph.

3.2.3. Community Analytics

Analytics on the detected communities help users gain better intuitions about them. Statistics per community provide summaries of each community's users. Word clouds and hashtag plots present the most frequently used hashtags and help users identify the most prominent topics per community. The frequencies of the most popular hashtags shared within a community are plotted. With regards to the posting activities of users, a time series plot shows the number of tweets shared per day by users of the community, revealing activity patterns.

Finally, a centrality/influence scatter plot is produced for the top 10 users of each community. The x axis shows the number of followers with respect to the number of followings, and the y axis corresponds to the ratio of incoming to outgoing interactions. Betweenness centrality is illustrated with bubbles; the larger the bubble, the higher the value of the feature calculated for a node. This plot helps identify community users that have essential roles in the spread of information (a high value of betweenness centrality) in correlation with their popularity rate (the x axis) and interaction rate (the y axis). The accounts in this plot are divided into four categories based on their positions. (i) *hot posts*: These have equal or smaller number of followers than followings and can be considered "regular" users (not popular). Their tweets have a strong influence on the spread of information as they have attracted the interest of other users. (ii) *Influencers*: These have higher numbers of followers than followings and can be considered popular. Their tweets have attracted the interest of other users, and their posts play vital roles in a community's topic and the spread of information. (iii) *Curators*: These have higher numbers of followers than followings and are regarded as popular. They have high posting activity levels as they usually post tweets and reference other accounts more than the opposite. Their beliefs are essential parts of a community's topic. (iv) *Unimportant*: These accounts have an equal or smaller number of followers than followings and are not popular. Their tweets do not attract other users' interest. Figure 4 presents an example of the analytics for a Twitter community. In the case of FB graphs, a heatmap of reactions per community shows the distribution of interactions on the posts of the top 10 Facebook pages/groups of the community based on the average number of total interactions per post.

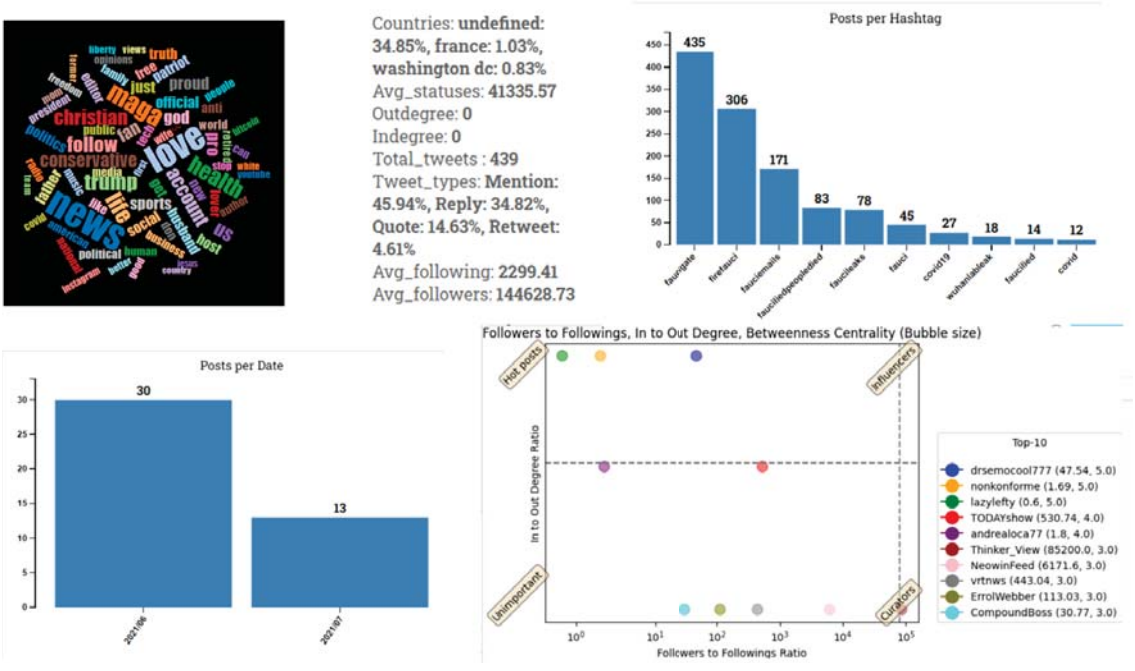


Figure 4. Community analytics: word cloud, user statistics, hashtag plot, tweet-posting plot, and centrality/influence scatter plot.

3.2.4. Propagation of URLs, Media, and Hashtags

To help study the spread of content in a network, we collected the top 10 URLs, media items, and hashtags in a network, and for each of them, we presented the interactions between the users. It is worth noting that there were items that appeared only inside a community and others that were disseminated across different communities.

3.2.5. Metagraph

The metagraph tool provides a higher-level representation of extracted communities and their interactions using a metagraph view: nodes correspond to communities, and their edges are weighted based on the interactions among the accounts of the respective communities. Wider edges between two communities mean a higher number of interactions across them. This view provides rich information about the relationships between the accounts of two and more communities. The metagraph view aims to reveal the degree to which the topics of different communities are related to each other. Wider edges indicate a higher correlation of topics.

3.2.6. Node and Edge Filters

Filters aim to offer a more targeted investigation and allow a user to limit information based on specific criteria. We grouped filters according to their types and placed them at the top left of the graph for easy navigation. Users with few interactions (a low degree) might not be of interest during an investigation and may therefore be removed using the min log-scaled degree filter. Additionally, if a user needs to investigate the interactions between accounts during a specific time interval, they can use the interactions date filter. Beyond node-based filters, there are also three types of edge filters depending on the type of edge: tweet edges, URL edges, and hashtag edges. For example, a tweet edge filter maintains (or removes) edges that are retweets, mentions, replies, and quotes. Users may also combine filters, which offers significant versatility during investigations.

3.2.7. Node Coloring

The node coloring tool exposes eight features characterizing an account's (node's) behavior. When a user selects one such feature, node colors are printed in a color scale from white (low) to dark blue (high) based on their values on this feature. These features include the following.

- `In-degree` quantifies the popularities of nodes, i.e., how much a node is referenced/linked to by others.
- `Out-degree` shows the extroversion of nodes. A node that references many other nodes on its tweets has a high value in this feature.
- `Out-degree to in-degree` is the ratio of out- to in-degree. Accounts that regularly reference other nodes in their tweets and are rarely referenced by others have high values in this feature. Accounts that are both extroverted and popular have low values.
- `Followers` is the number of account followers.
- `Followings` is the number of accounts that the account follows.
- `Followings to followers` is the ratio of the number of followings to the number of followers. This feature quantifies how popular an account is (low value), how selective it is with its followings (low value), and how likely it is to follow back (high value).
- `Betweenness centrality` captures the role of a node in the spread of information across a network. Higher values indicate more important roles.
- `Similarity of tweets` shows how similar the posts of an account are. A node with a high value in this feature regularly posts similar content on its tweets.

3.2.8. Highlight Suspicious Accounts

Finally, we implemented six features that, when combined with three of the above, indicate suspicious accounts spamming or spreading disinformation. Inspired by features presented in the literature to train machine-learning models and detect spam posts [14], we implemented a set of features that support the interactive exploration of a dataset to find accounts that were worth further investigation. The features were normalized using min-max normalization. To calculate extreme values on these features and subsequently highlight accounts with such values, we used quartiles and boxplots. The accounts were highlighted in red on the graph, providing a semi-automatic identification of suspicious accounts. The implemented features are listed below.

- `Following rate` is the ratio of the number of followings to the number of days since an account was first created.
- `Status rate` is the ratio of the number of posts to the number of days since an account was created.
- `Average mentions per post` shows the average number of mentions in an account's tweets. A common strategy for spreading disinformation is mentioning many accounts in tweets.
- `Average mentions per word` shows the average number of mentions in a tweet's text. The tactic of posting tweets with many mentions and a single hashtag is often regarded as spam-like or suspicious. This feature is normalized to the total number of posts.
- `Average hashtags per word` calculates the average number of hashtags in a tweet's text.
- `Average URLs per word` calculates the average number of URLs in a tweet's text.

Figure 5 illustrates an example of a Twitter graph for a Fauci use case. At the left of the figure, the betweenness centrality node coloring is applied to highlight nodes with high influence over the flow of information in the graph and that are worth investigating. At the right of the figure, all highlights of suspicious account features are selected, and 425 accounts are highlighted as suspicious, limiting users who are worth evaluating and offering a user a clue of where to start an investigation.

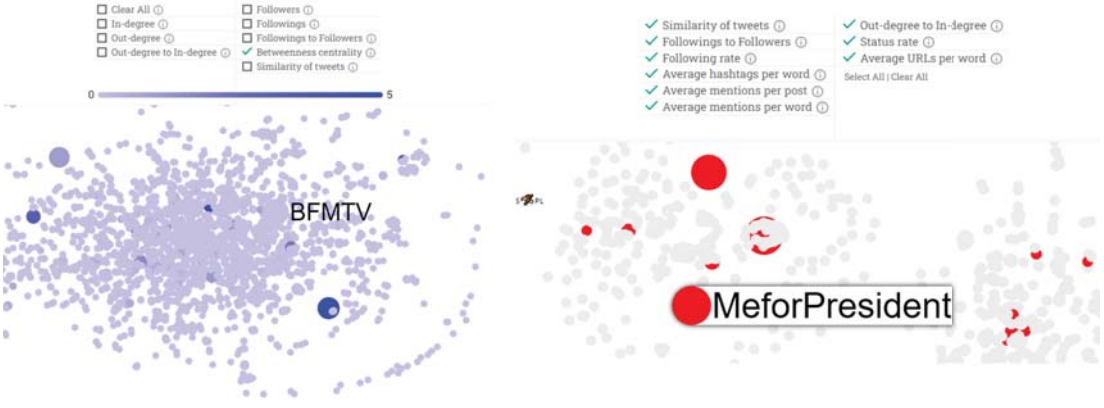


Figure 5. Example of the node-coloring filter (**left**) and highlighting suspicious accounts filter (**right**).

Table 4 summarizes the features/functionalities supported per platform. Although we aimed to adapt all the developed features on all three platforms, the available data restricted the implementation of features in some cases.

Table 4. List of developed features and the platforms that are supported.

Features/Functionalities	Twitter	Facebook	Telegram
Communities	✓	✓	✓
Individual account inspections	✓	✓	✓
Post text	✓	✗	✗
Statistics per community	✓	✓	✗
Word clouds per community	✓	✓	✓
Centrality plots per community	✓	✗	✗
Date plots per community	✓	✓	✓
Hashtag plots per community	✓	✗	✓
Heatmaps of reactions per community	✗	✓	✗
Propagation flow of top 10 URLs	✓	✓	✓
Propagation flow of top 10 media	✓	✓	✓
Propagation flow of top 10 hashtags	✓	✗	✓
Metagraphs	✓	✓	✓
Edge Filters	✓	✗	✗
Node Filters	✓	✓	✓
Node coloring	✓	✓	✓
Suspicious accounts	✓	✗	✗

4. COVID-19-Related Use Cases

4.1. Twitter Data

We collected four COVID-19-related datasets for topics for which disinformation was prominent. To collect the datasets, we used the Twitter search API, querying with the hashtags #FireFauci, #FauciGate, #Hydroxychloroquine, #BigPharma, and #GreatReset. We collected all tweets containing these hashtags posted between 1 June 2021 and 15 July 2021. Table 5 presents the dataset statistics.

Table 5. Statistics for the collected COVID-19-related Twitter disinformation datasets.

	Fauci	Hydroxychloroquine	Big Pharma	Great Reset
Total tweets	18,500	6239	16,568	13,380
Retweets	4790	3597	9667	6780
Quotes	7787	1114	2281	2615
Tweets with replies	4696	1046	3579	3037
Tweets with mentions	4926	2439	5609	4884
User-posted tweets	11,155	4310	10,474	8175
Total users in tweets	18,310	7078	18,175	14,716

The Fauci and hydroxychloroquine cases came after the Washington Post and BuzzFeed News filed Freedom of Information Act requests for Dr. Anthony Fauci's emails, published that correspondence on 1 June 2021, and showed how Dr. Anthony Fauci navigated the early days of the COVID-19 pandemic. The emails contained discussions about what Fauci was told on the origins of the coronavirus, what he knew about the drug hydroxychloroquine, and what he said about the use of face masks. Apart from informing the public, this email disclosure led to the propagation of misleading facts about COVID-19 by decontextualizing parts of the discussions. Fauci's political opponents and several conspiracy theorists took the opportunity to spread their beliefs on social networks by sharing out-of-context claims.

Another conspiracy theory that gained popularity was the Big Pharma theory. A group of conspiracy theorists claimed that pharmaceutical companies operate for sinister purposes and against the public good. They claimed that the companies conceal effective treatments or even cause and worsen a wide range of diseases.

Finally, the Great Reset theory referred to a theory that the global elites have a plan to instate a communist world order by abolishing private property while using COVID-19 to solve overpopulation and enslave what remains of humanity with vaccines.

4.2. Facebook Data

We used CrowdTangle to collect Facebook posts on the aforementioned COVID-19-related use cases. We submitted four search queries in CrowdTangle using the hashtags #FireFauci, #FauciGate, #Hydroxychloroquine, #BigPharma, and #GreatReset as search keywords. The search retrieved Facebook posts by public Facebook pages and groups (but not posts by Facebook users due to Facebook graph API limitations). We ended up with two datasets, Fauci and hydroxychloroquine. The Big Pharma and Great Reset topics were discarded due to very low numbers of retrieved posts. Table 6 lists the statistics for the Fauci and hydro datasets.

Table 6. Statistics for the collected COVID-19-related Facebook disinformation datasets.

	Fauci	Hydroxychloroquine
FB posts	553	1572
FB groups/pages	352	984
Articles	95	504
Photos	109	264
Videos	71	53

4.3. Telegram Data

In order to acquire information from public Telegram channels, a specific data-acquisition system was crafted. A core component of such a system is, usually, Web Scraper, the aim of which is to parse HTML markup code and pass it through a set of selectors in order to structure the acquired information made available via a Web user interface. A relevant difference between the platform Telegram and other platforms, such as Facebook and Twitter, is the absence of an internal content-search feature. Thus, only public channels and groups can be found via

a global search feature—not the messages contained in them. In order to find public channels of interest, two different approaches were followed. The first was executing keyword-based searches via a Google custom search engine specifically designed for Telegram content (https://cse.google.com/cse?cx=004805129374225513871:p8lhfo0g3hg accessed on 8 April 2022). The second approach was running hashtag-based searches on Twitter and applying a filter to them in order to receive only results containing at least one URL referencing Telegram content. We chose to select up to three public channels per each hashtag. Specifically, we selected from the first three top tweets, (https://help.twitter.com/en/using-twitter/top-search-results-faqs accessed on 8 April 2022) which Twitter valued as most relevant (in descending order) at the time when we executed the search. For both approaches, we decided to select only the first three results provided in order both to keep the information consistent and to have a sufficiently sized dataset. Subsequently, the information populating our Telegram dataset was acquired from chats by the following identified handles: @RealMarjorieGreene, @QtimeNetwork, @cjtruth316, @trumpintel, @Canale_Veleno, @WeTheMedia, @shaanfoundation, @TerrenceK-Williams, and @zelenkoprotocol. Specifically, the first three were selected for the #*FireFauci* dataset, the second three were selected for the #*FauciGate* dataset, and the last three were selected for the #*Hydroxychroroquine* dataset. Table 7 lists the statistics for the #FireFauci, #FauciGate, and #Hydroxychloroquine datasets.

Table 7. Statistics for the collected COVID-19-related Telegram disinformation datasets.

	FireFauci	FauciGate	Hydroxychloroquine
Subscribers	326,586	6488	186,236
Messages	14,762	181,700	13,422
URLs	6453	83,993	10,032
Hashtags	871	18,653	106

4.4. Iterative Evaluation and Feedback

We applied the analysis to the four collected COVID-19-related use cases, which are cases of significant value that reflect the challenge of disinformation. These use cases arose in the context of user-driven evaluation activities that took place within the Horizon 2020 WeVerify project (https://weverify.eu/ accessed on 8 April 2022). Journalists and fact-checkers participated in these evaluation activities and provided helpful feedback on the proposed MeVer NetworkX analysis and visualization tool. The users received brief guidelines on the functionalities of the tool and query files on a set of use cases collected within the project (beyond the four use cases presented here). They analyzed the query cases and provided comments/suggestions on the parts of the tool that were unclear to them and parts that would make the analysis easier to digest and more efficient. We enhanced the tool with the user feedback and came up with the final version that is presented in this paper.

5. Analysis Using the MeVer NetworkX Analysis and Visualization Tool

The main focus of our analysis was to simulate a scenario in which, through the tool, an end user tries to identify and inspect suspicious accounts within a given dataset graph.

5.1. Fauci

5.1.1. Twitter Analysis

The graph included 18,310 nodes and 27,882 edges. Different colors were assigned to the nodes of different communities. We first selected the all option of the suspicious account filter at the top of the graph. The resulting 425 accounts highlighted as suspicious were presented in the graph in red, as shown in Figure 6. We queried the Twitter search API three months after the dataset was collected, and 78 of the 425 likely suspicious accounts did not exist on Twitter anymore due to violating Twitter policies. This indicates that the tool likely correctly highlighted at least 78 accounts. Note that the fact that the remaining

347 accounts were still active 3 months after the dataset collection does not mean that they were not suspicious.

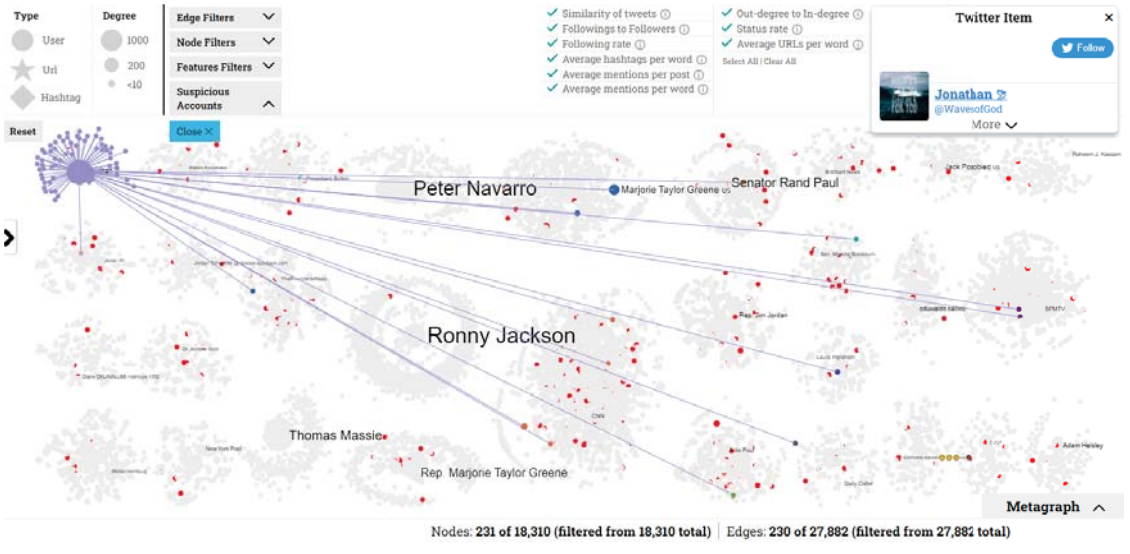


Figure 6. Example of a suspicious account in the Fauci analysis.

For instance, an account that was flagged as suspicious by the tool was the account with the highest number of interactions, (*@WavesofGo*), in Community 14 (Figure 6). The account had 228 interactions and was the largest circle in the community. The interactions corresponded to mentions of 228 individual accounts in the tweets, while there were no interactions toward this account from other users. The mentions referred to popular accounts, such as President Biden's *@POTUS* official account and Marjorie Taylor Greene's *@mtgreenee* account. Additionally, the similarity of tweets feature indicated that the user posted the exact tweet text each time by referencing different accounts. Table 8 shows two examples of tweets shared by this account through the network, indicating that the account was a strong supporter of Christianity and an opponent of vaccination. Twitter suspended the account due to not complying with Twitter's policies. We then inspected the statistics on Community 14. The hashtag plot revealed the community's main topics were (#JesusSaves and #AntiVax), as shown in Figure 7a.

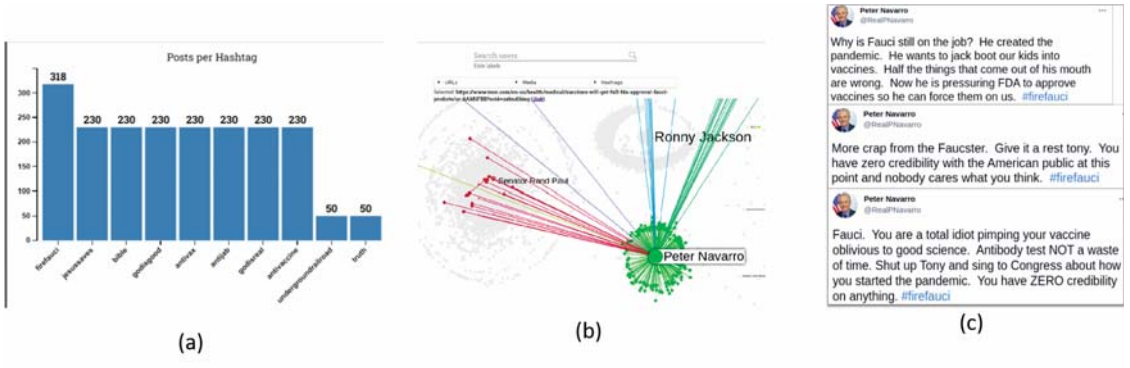


Figure 7. Results of Fauci use-case analysis: (**a**) hashtag plots of Community 14, (**b**) an example of MSN article spread, and (**c**) Peter Navarro's tweets against Fauci.

Table 8. Tweets of *@WavesofGo* account that indicate the account was a strong supporter of Christianity and was against vaccination.

"Jesus heals the sick when men had leprosy, the worst diseases back then and people were afraid to be around them. Don't take Covid Vaccine because JESUS IS A HEALER. Repent and confess to Jesus. #AntiVax #AntiVaccine #AntiJab #GodIsReal #FireFauci #Bible #GodisGood #JesusSaves"	"Praise God! YES Jesus heals the sick and did so when men had leprosy and people were afraid to be around them. That's why Christians shouldn't take Covid Vaccine because our GOD IS A HEALER. #AntiVax #AntiVaccine #AntiJab #GodIsReal #FireFauci #Bible #GodisGood #JesusSaves #God"

Similarly, the most active account of Community 10, *Adam Helsley (@AdamHelsley1)*, was flagged as suspicious. The account only replied to other accounts with the #FireFauci hashtag, trying to campaign against Fauci. The account remained active on Twitter three months after the data collection since its behavior complied with Twitter's policies, even though its posting activity resembled that of a spammer. Our tool highlighted this as suspicious behavior. The end user is responsible for investigating further and deciding whether this account tries to adversely affect other users. This account triggered five of the nine suspicious features, namely the ratio of out-degree to in-degree, the average number of mentions per post and word, the average number of hashtags per word, and the similarity of tweet text features.

For a further investigation, we used the Botometer, a tool that checks the activity of a Twitter account and gives it a score on how likely the account is a bot. The Botometer's values range between zero and five, and the higher the value, the more likely the account is a bot. Despite the above signs, the Botometer did not classify the account as a bot.

Next, we investigated the propagation flow of the top 10 URLs, hashtags, and media items. By selecting an item, a user can monitor the item's dissemination across a network. In this case, one of the most popular URLs in the network was an *msn.com* article (https://www.msn.com/en-us/health/medical/vaccines-will-get-full-fda-approval-fauci-predicts/ar-AAM1FBB?ocid=uxbndlbing accessed on 8 April 2022), which says that Dr. Fauci advises vaccinated Americans to wear masks in areas with low COVID-19 immunization rates. This topic attracted Twitter users' interest and indicated that it was worth a further investigation. This article was mainly spread in Community 3 (Figure 7b). The account of this community interacting the most frequently was *Peter Navarro (@RealPNavarro)*. Peter Kent Navarro is an American economist and author who served in the Trump administration as assistant to the president, director of trade and manufacturing policy, and policy coordinator of the National Defense Production Act. This account has more than 161 thousands followers. He posted aggressive tweets against Fauci that spawned heated discussions on Twitter. From his 1373 total interactions, only five referenced other accounts, while in the remaining 1368, he was referenced by other accounts. Figure 7c shows his top three tweets that gained the most interactions in the network and reveal his negative attitude toward Fauci.

Finally, the word cloud plot of Community 3 provided insights into the topics of the accounts that made it up. In Figure 8a, we observe that words such as "Trump" and "patriot" frequently appeared, concluding that many accounts in this community are likely Trump supporters and explaining the reason for the attack against Fauci. Moreover, the centrality plot, shown in Figure 8b, labels Peter Navarro as an influencer.

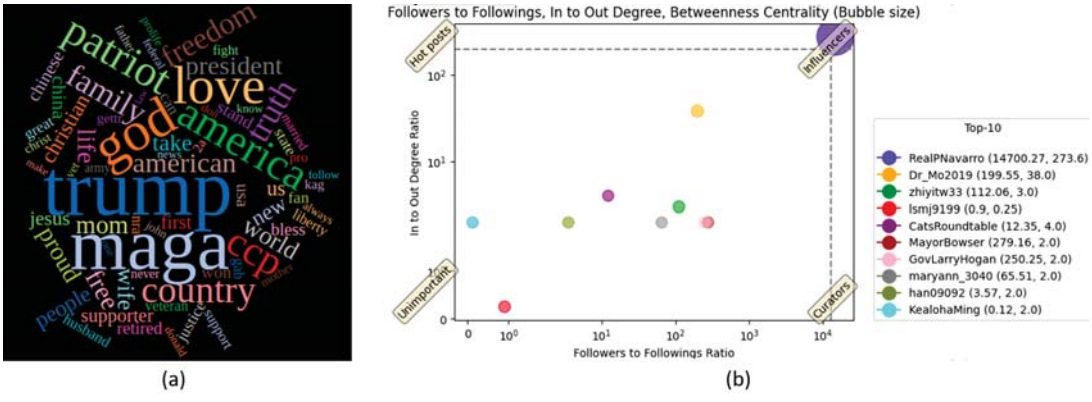


Figure 8. Results of Fauci use-case analysis: (**a**) word cloud of Community 3 and (**b**) centrality plot of Community 3 indicating that Peter Navarro's account stands as an influencer in the community.

5.1.2. Facebook Analysis

The 627 nodes and 553 edges derived from the Facebook Fauci dataset formed 29 communities. In all, 19 of the 29 communities consisted of three nodes, which we discarded as unimportant; we focused on the remaining 10 communities but noticed that the information diffusion among these communities was limited. Figure 9 illustrates the corresponding metagraph. We noticed that the largest community (with the ID 0) had no interactions with other communities at all.

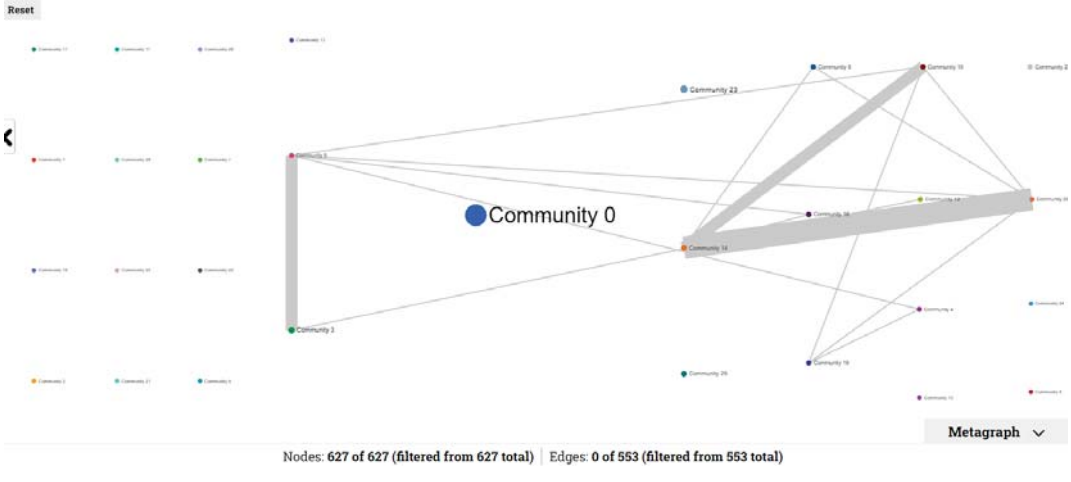


Figure 9. Metagraph showing that the information diffusion in the network was limited.

The node most frequently interacted with in Community 22 was an article from the Convention of States Action, which was the most-shared link in the network. The article references a poll about the aspect of parents with respect to the vaccination of their children. It states that Dr. Fauci lied to people, which led to hesitation among parents about their children's vaccination. The article was shared among the 52 nodes of Community 22, creating a network of supporters.

Using the tool, we managed to draw some interesting observations that provided insights about the network users' attitudes toward the topic. The positive reactions filter highlighted nodes that received love, wow, haha, and care reactions. Figure 10 shows the

highlighted nodes of Community 8 that were mainly around Marjorie Taylor Greene, a Republican congresswoman who sparked the #FireFauci hashtag on Twitter. The network revealed positive emotions around Marjorie Taylor Greene, which was the opposite of the aggressive attitudes toward Fauci. Additionally, through the propagation flow feature, a link of the top 10 shared links in the network pointed to a Facebook image post containing a screenshot of a Marjorie Taylor Greene post urging for Fauci's dismissal.

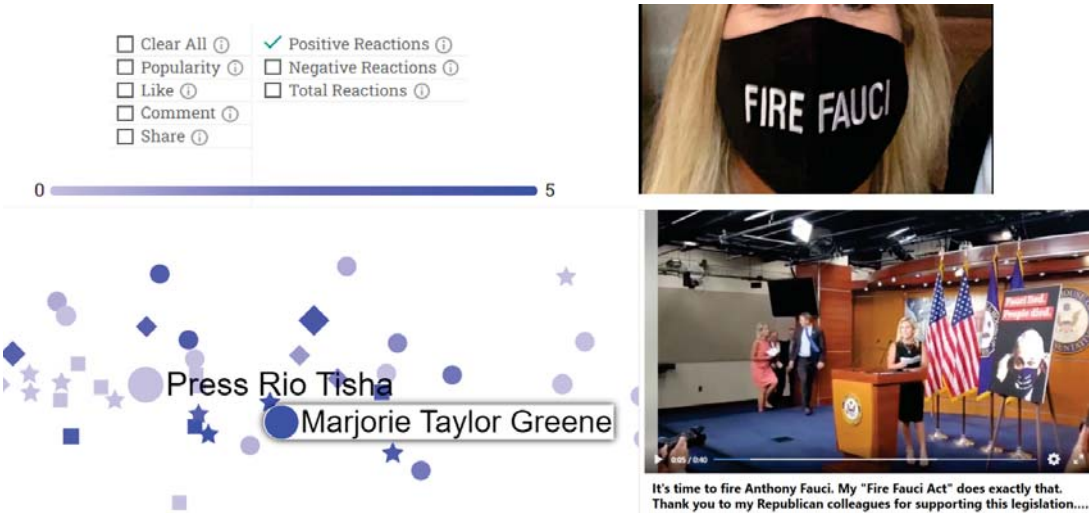


Figure 10. Inspecting the positive reactions filter feature and retrieving the popular account of Marjorie Taylor Greene, which aims to influence the public against Fauci.

The tool's search bar allows users to cross-check whether the same user appears in Twitter and Facebook graphs. Although Marjorie Taylor Greene is significantly involved in information dissemination in the Facebook graph, her account does not appear in the Twitter graph. However, Senator Rand Raul, the most famous account in the Twitter graph, is also active in the Facebook graph. Specifically, an image post with text "I told you so" was shared, and it gained a lot of engagement (approx. 2000 comments and 8000 shares).

5.1.3. Telegram Analysis

For the analyses of the #FauciGate and #FireFauci use cases, we filtered the collected Telegram data with a time interval of between 1 June 2021 and 15 July 2021. We discovered 985 nodes and 1043 edges, which formulated a graph with 15 communities. The largest community consisted of 381 nodes, while there were nine communities with one node. The word clouds of the largest communities in the graph are illustrated in Figure 11. Apart from Community 6, the rest of the communities' word clouds showed a wide range of topics discussed within the Telegram channels and no specific focus on the Fauci case. This made it more challenging to collect data from Telegram and analyze a particular topic of interest. From the propagation flow feature, it could be seen that half of the top 10 URLs in the graph pointed to Telegram messages or accounts, indicating that the discussions in the Telegram channels stayed within Telegram. The top URL was a YouTube link of a video that had already been removed at the time of the analysis because it violated YouTube's community guidelines. This YouTube video appeared in Community 6, in which @RealMarjorieGreene was the most interactive node. @RealMarjorieGreene is an account that seemed to be involved in the Fauci case during the analyses of the Twitter and Facebook graphs.

Figure 11. Telegram word clouds.

We concluded that the analysis and visualization of narratives discussed within Telegram are currently limited. To provide helpful insights and highlight suspicious accounts or behaviors, we need to focus on the data collection step—the limited data acquired by Telegram results in shallow analyses.

5.2. Hydroxychloroquine

A similar procedure was followed for the analysis of the hydroxychloroquine use case. The tool detected 110 out of 7078 accounts as suspicious and likely propagating disinformation; 15 of these 110 accounts do not exist on Twitter anymore.

A suspicious account was found in Community 20 with the name *@nyp64N5uCEe3wiu*; it was suspended from Twitter for violating Twitter's rules and policies. Within the network, the account interacted with 27 other accounts in total, and 26 of these interactions replied to the tweets of other accounts or mentioned other accounts. The account acted like a spammer by posting the same text—only hashtags—which was highlighted by the similarity of tweets feature. The hashtags that the account was disseminating included #CCPVirus, #TakeDowntheCCP, #DrLiMengYan, #Hydroxychloroquine, #GTV, #GNews, #NFSC, #WhistleBlowerMovement, #LUDEMedia, #UnrestrictedBioweapon, #COVIDVaccine, #COVID19, #IndiaFightsCOVID, #OriginofCOVID19, #Coronavirus, #WuhanLab, #CCP_is_Terrorist, #CCPisNotChinese, #CCPLiedPeopleDied, and #MilesGuo. Within Community 20, there were also eight accounts that were highlighted as suspicious creating, a doubt about Community 20 as a whole. A user could further investigate each of these accounts individually and draw some conclusions about their participation (or not) in the dissemination of information/disinformation.

5.3. Big Pharma Use Case

The tool labeled 220 out of 18,175 accounts as suspicious for the Big Pharma analysis. In all, 54 of these 220 suspicious accounts do not exist on Twitter anymore. An example of a suspicious account was *@UnRapporteur1* (Figure 12), which was the most frequently interacting account of Community 24 (visualized as the larger circle). The account posted the exact same text in its tweets by referencing other accounts. Figure 12 presents the text of the tweets, which contained offensive language. This account is still active on Twitter and is likely suspicious with high values for four features out of nine: the ratio of out-degree to in-degree, the average number of mentions per post and word, and the similarity of tweets' text. Botometer rated this account as a bot with high confidence (5/5).

Another suspicious account that was worth investigating was *@checrai71*, a member of Community 2. The tool flagged this account with the ratio of out-degree to in-degree, the average number of mentions per post and per word, and the followings to followers ratio. The account posted 29 tweets mentioning and replying to 74 different accounts. This account is still active on Twitter and has proven to be a strong supporter of the Big Pharma conspiracy theory based on a video (https://vimeo.com/500025377 accessed on 8 April 2022) shared in its tweets supporting this theory. The Botometer's score of this account was 2.2 (i.e., leaning more toward a "regular user" rather than bot).

Figure 12. Example of a suspicious account in the Big Pharma analysis (**right**) and example of a suspicious account posting conspiracy tweets in the great reset analysis (**left**).

5.4. Great Reset Use Case

Regarding the great-reset-related tweets, the tool highlighted 231 out of 14,716 accounts as suspicious. From the highlighted accounts, we found that 29 are not available on Twitter anymore, and we further inspected one of them, which was *@riesdejager*. This account tweeted the conspiracy message presented in Figure 12. It was labeled as having two suspicious features: similarity of tweet texts and average number of URLs per word.

The account with the highest number of suspicious features was *@inbusiness4good*. It was highlighted due to its ratio of out-degree to in-degree, average number of mentions per post, similarity of tweets texts, average number of URLs per word, and average number of hashtags per word. It supported an action called Clim8 Savers, and the purpose was to persuade people to plant trees. The Botometer's score of this account was 3.4.

5.5. Quantitative Analysis of Suspicious Users

We further analyzed the developed features that highlighted the users as suspicious and investigated the importance of each feature in the four use cases. As presented in Table 9, the cosine similarity of tweet feature dominated in the hydro and Big Pharma cases. In contrast, the average mentions per post feature was the top feature in the Fauci and great reset cases. Notably, there were cases in which a feature was not involved but dominated in other cases. For example, the average mentions per word feature did not highlight any user in the hydro and great reset cases; however, in the Big Pharma case, this feature labeled 50 out of the 220 likely suspicious users (22.7%). A manual inspection of the four COVID-19-related use cases concluded that it was more likely a user was spreading disinformation when more features highlighted the user. However, there were cases in which even one feature was a solid indication for a further investigation of the user.

Table 9. The number of users that each feature highlighted as a suspicious per use case.

Suspicious Account Features	Fauci	Hydro	Big Pharma	Great Reset
Out-degree to in-degree	62	3	18	10
Followings to followers	28	7	18	13
Following rate	27	14	33	35
Status rate	19	6	16	16
Average mentions per post	**143**	0	61	**110**
Average mentions per word	8	0	50	0
Average hashtags per word	18	37	1	4
Average URLs per word	57	9	11	11
Cosine similarity of tweets	122	**44**	**62**	50

Table 10 presents the number of users per investigated case and the percentages of nonexistent, suspended, and suspicious users (as labeled by our tool). Additionally, at the bottom of the table, the percentage of users calculated with bot scores (https://botometer.osome.iu.edu/ accessed on 8 April 2022) higher than three are listed.

It is noticeable that the MeVer NetworkX analysis and visualization tool labeled each user with a low percentage (approx. 1–2%) of the total number of users in a network as suspicious. In that way, the tool provided users with a clue to start an investigation and focus on a few users with one or more features that raised suspicions. Using the Twitter API two months after we collected the datasets, we found that out of the detected suspicious users, 10.1% for the Fauci case, 13.6% for the hydro case, 8.6% for the Big Pharma case, and 12.5% for the great reset case were not available on Twitter anymore. In this way, we can consider the highlighted users as likely correct selections (i.e., users that violated Twitter's policies). However, users who remained active but were still highlighted by the tool could spread disinformation or support misleading claims, but Twitter's policies were not violated. Such an example is the account @UnRapporteur1, described in Section 5.4.

Table 10. Quantitative analysis of suspicious users for the four COVID-19-related use cases.

	Fauci	**Hydro**	**Big Pharma**	**Great Reset**
All users	18,310	7078	18,175	14,716
Nonexistent	1529	477	1261	1187
Suspended	868	257	653	610
Suspicious	425	110	220	231
Percentage of suspicious users in relation to all users	2.3%	1.6%	1.2%	1.6%
Percentage of suspicious users not available on Twitter	10.1%	13.6%	8.6%	12.5%
Percentage of suspicious users not available on Twitter due to being suspended	6.1%	6.4%	5.5%	5.6%
Users with bot scores (not available for unavailable users)	16,635	6548	16,755	13,396
Percentage of users with bot scores ≥ 4	10.7%	6.8%	7.0%	7.12%
Percentage of users with 4 >bot scores ≥ 3	13.9%	21.2%	19.3%	16.1%

6. Execution Time

To study the computational behavior of the proposed tool, we collected a large Twitter dataset by querying with the hashtag #COVIDVaccine. The hashtag was selected as it was a trending topic, which could result in large networks. We generated graphs of progressively larger sizes (sums of nodes and edges), starting from ∼1000 and reaching up to ∼140,000, which we considered sufficient for the support of several real-world investigations. We carried out a graph analysis and visualization on a Linux machine with a 2 Intel Xeon E5-2620 v2 and 128 GB of RAMand calculated the time needed for each graph. Figure 13 illustrates the execution times in seconds.

We noticed that for the small graphs (fewer than 10,000 nodes and edges), the execution time increased linearly by a very small factor in relation to the graphs' sizes. For instance, doubling the number of nodes and edges from 700 to 1400 nodes and edges resulted in an increase of 12.5% in execution time. For larger graphs, the execution time increased with a much higher linear factor or even in a super-linear manner. The time needed to analyze and visualize a graph with a size of 140,000 nodes and edges was twice what it took to build a graph with a size of ∼72,000 nodes and edges.

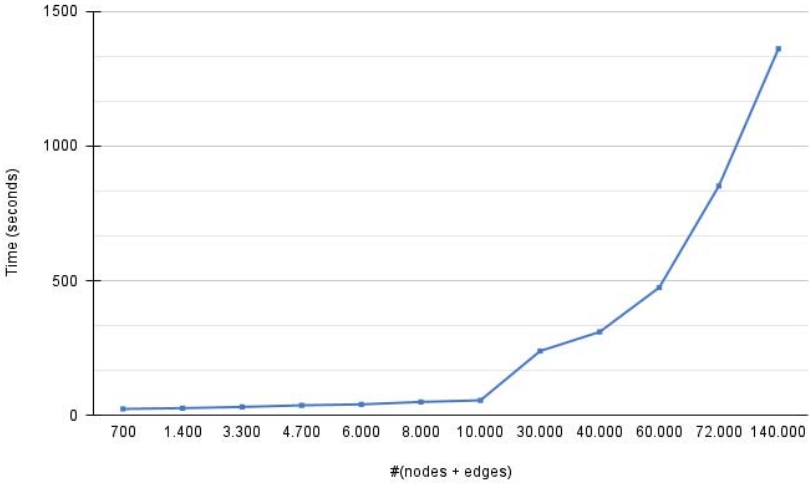


Figure 13. Execution times in seconds needed to build and visualize small and large graphs.

7. Comparison with Gephi and Hoaxy

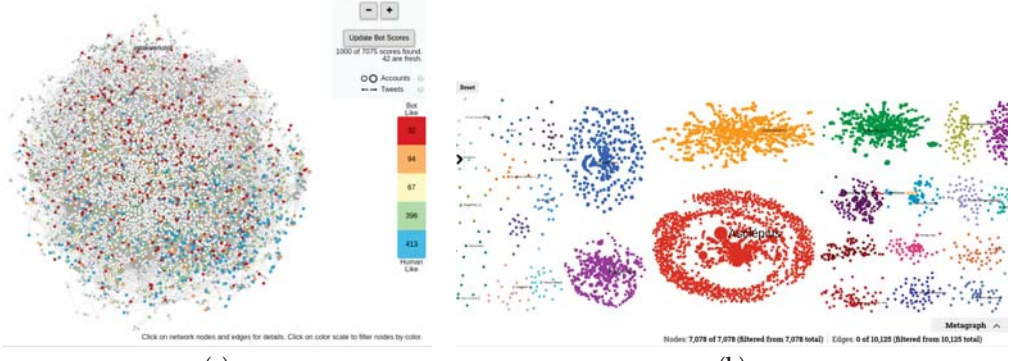
Gephi is one of the top visualization tools in social network analysis [30,31]. It is a free open-source standalone software used for visualizations and explorations of all kinds of networks. Its advantages include high quality visualizations and the fact that knowledge of programming languages and the ability to handle large graphs are not required. Although Gephi is a leading tool for network visualizations, our MeVer NetworkX analysis and visualization tool is not comparable to Gephi. Gephi is a standalone software, as aforementioned, while the MeVer NetworkX analysis and visualization tool is provided as a web-based application. Gephi provides visualizations for large networks (i.e., it can render networks up to 300,000 nodes and 1,000,000 edges), while our tool supports smaller graphs, focusing on a specific narratives of disinformation. The main difference between the tools that makes them incomparable is that Gephi provides a multitude of functionalities for visualization and filtering, while our tool focuses more on the analysis of the accounts involved in a network, their characteristics, and the information that is disseminated through and among them.

Hoaxy is a tool that visualizes the spread of information on Twitter. It supports the uploading of a CSV or JSON file containing Twitter data. For a comparison, we created CSV files compatible with Hoaxy containing the tweets of the four COVID-19-related use cases that we investigated. We submitted each file and created the graphs with Hoaxy. First, we examined the execution time needed to analyze and build the graphs. In Table 11, Hoaxy seems much faster than MeVer NetworkX in all use cases. However, we needed to consider each tool's features to decide which one is faster. Based on this, we created Table 12, in which the features of the two tools are presented side by side. Hoaxy provides far fewer features than the proposed tool. Concerning execution time, Hoaxy is faster in terms of the time needed to analyze input data and build a graph.

Table 11. Execution times of MeVer NetworkX analysis and visuazalion tool vs. Hoaxy for the four COVID-19-related use cases.

	Fauci	**Hydro**	**Big Pharma**	**Great Reset**
Hoaxy (time in s)	240	105	230	185
MeVer (time in s)	355	139	349	261

Apart from the multitude of features that the MeVer NetworkX tool provides over Hoaxy and its comparable execution time, a significant advantage of the MeVer NetworkX tool is the improved layout with non-overlapping graphs, providing users with an easy-to-digest visualization of communities. Figure 14 illustrates the graphs around the hydro topic in MeVer NetworkX (right) and Hoaxy (left). In the graph built by Hoaxy, the nodes and edges overlap and a user must zoom in and out to investigate them. Instead, the MeVer NetworkX tool provides a simple and clear visualization with different colors among communities. Moreover, the node with its interactions is highlighted by clicking on a node while the rest become blurred. In this way, a user can more easily inspectthe nodes of interest one by one.



(**a**) (**b**)

Figure 14. Graphs for hydro use case: (**a**) Hoaxy visualization and (**b**) MeVer NetworkX analysis and visualization.

Table 12. Comparison of MeVer NetworkX's and Hoaxy's analysis and visualization features.

Feature	MeVer Tool	Hoaxy	Feature	MeVer Tool	Hoaxy
Community detection	+	–	Centrality/influence scatter plots	+	–
Individual user inspections	+	+	Propagation flows of URLs	+	+
Tweet texts	Embedded	External links	Propagation flows of media	+	–
Word clouds	+	–	Propagation flows of hashtags	+	–
Statistics for each community	+	–	Metagraphs	+	–
Hashtag plots	+	–	Tweet timelines	+	+
Date plots	+	–	Highlight of suspicious accounts	+	–

8. Discussion and Future Steps

The tool is available upon request (https://networkx.iti.gr/ accessed on 8 April 2022). To the best of our knowledge, it is the only tool supporting the analysis of multiple platforms and even providing some cross-platform investigations. The tool aims to support the demanding work of journalists and fact checkers to combat disinformation. The advanced functionalities offered by the tool are valuable, as showcased through the presented use cases. The aggregation and visualization capabilities provided to users offer easy ways to navigate large graphs without a need for special knowledge. The developed functionalities offer users a semi-automatic procedure that can increase productivity and save time. The tool's core functionality is to highlight suspicious accounts based on features that are often associated with inauthentic behavior. Although the presented use cases showed that these features are helpful and provide valuable insights about the accounts, in the future, we aim to train models that automatically highlight suspicious users, providing better support to investigators. Additionally, a direction that we are investigating as a future step is integrating third-party tools, such as the Botometer, which provide more insights about

accounts. Finally, a data-collection component is an essential part of the tool. The tool requires GEFX or CSV files containing data collected by a social media platform in question. However, this part of collecting data needs specialized knowledge or some third-party tools. For that reason, we integrated the InVID-WeVerify plugin into the tool in order to offer a smooth and intuitive analysis process and are considering further ways to improve the user experience.

Author Contributions: Conceptualization, O.P., T.M. and S.P.; Data curation, O.P., T.M. and F.P.; Formal analysis, O.P. and S.P.; Funding acquisition, S.P. and I.K.; Investigation, O.P.; Methodology, O.P., T.M. and S.P.; Project administration, S.P.; Software, O.P. and L.A.; Supervision, S.P.; Validation, O.P., T.M. and S.P.; Visualization, O.P.; Writing—original draft, O.P. and F.P.; Writing—review & editing, O.P. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the WeVerify and AI4Media projects, which are funded by the European Commission under contract numbers 825297 and 951911, respectively, as well as the US Paris Tech Challenge Award, which is funded by the US Department of State Global Engagement Center under contract number SGECPD18CA0024.

Data Availability Statement: Not Applicable, the study does not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Posetti, J. News industry transformation: Digital technology, social platforms and the spread of misinformation and disinformation. In *Journalism,'Fake News' and Disinformation: A Handbook for Journalism Education and Training*; UNESCO: Paris, France, 2018. Available online: https://bit.ly/2XLRRlA (accessed on 8 April 2022).
2. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
3. Meserole, C. *How Misinformation Spreads on Social Media—And What to Do about It*; The Brookings Institution: Washington, DC, USA, 2018. Available online: https://www.brookings.edu/blog/order-from-chaos/2018/05/09/how-misinformation-spreads-on-social-media-and-what-to-do-about-it (accessed on 8 April 2022).
4. Himelein-Wachowiak, M.; Giorgi, S.; Devoto, A.; Rahman, M.; Ungar, L.; Schwartz, H.A.; Epstein, D.H.; Leggio, L.; Curtis, B. Bots and Misinformation Spread on Social Media: Implications for COVID-19. *J. Med. Internet Res.* **2021**, *23*, e26933. [CrossRef] [PubMed]
5. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.C.; Flammini, A.; Menczer, F. The spread of low-credibility content by social bots. *Nat. Commun.* **2018**, *9*, 1–9. [CrossRef] [PubMed]
6. Erik Borra, B.R. Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib J. Inf. Manag.* **2014**, *66*, 262–278. [CrossRef]
7. Marinova, Z.; Spangenberg, J.; Teyssou, D.; Papadopoulos, S.; Sarris, N.; Alaphilippe, A.; Bontcheva, K. Weverify: Wider and enhanced verification for you project overview and tools. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–4.
8. Teyssou, D.; Leung, J.M.; Apostolidis, E.; Apostolidis, K.; Papadopoulos, S.; Zampoglou, M.; Papadopoulou, O.; Mezaris, V. The InVID plug-in: Web video verification on the browser. In Proceedings of the First International Workshop on Multimedia Verification, Mountain View, CA, USA, 27 October 2017; pp. 23–30.
9. Peeters, S.; Hagen, S. The 4CAT capture and analysis toolkit: A modular tool for transparent and traceable social media research. *SSRN* **2021**, 3914892. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3914892 (accessed on 8 April 2022). [CrossRef]
10. Wang, B.; Zubiaga, A.; Liakata, M.; Procter, R. Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter. *arXiv* **2015**, arXiv:1503.07405.
11. ElAzab, A. Fake Account Detection in Twitter Based on Minimum Weighted Feature set. *World Acad. Sci. Eng. Technol. Int. J. Comput. Inf. Eng.* **2016**, *10*, 13–18.
12. Mateen, M.; Iqbal, M.A.; Aleem, M.; Islam, M.A. A hybrid approach for spam detection for Twitter. In Proceedings of the 2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 10–14 January 2017. [CrossRef]
13. Yang, C.; Harkreader, R.; Gu, G. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1280–1293. [CrossRef]
14. Rodríguez-Ruiz, J.; Mata-Sánchez, J.I.; Monroy, R.; Loyola-González, O.; López-Cuevas, A. A one-class classification approach for bot detection on Twitter. *Comput. Secur.* **2020**, *91*, 101715. [CrossRef]
15. Zhang, Z.; Hou, R.; Yang, J. Detection of Social Network Spam Based on Improved Extreme Learning Machine. *IEEE Access* **2020**, *8*, 112003–112014. [CrossRef]

16. Alom, Z.; Carminati, B.; Ferrari, E. Detecting Spam Accounts on Twitter. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018. [CrossRef]
17. Keller, F.B.; Schoch, D.; Stier, S.; Yang, J. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Commun.* **2019**, *37*, 256–280. [CrossRef]
18. El-Mawass, N.; Honeine, P.; Vercouter, L. SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields. *Inf. Process. Manag.* **2020**, *57*, 102317. [CrossRef]
19. Masood, F.; Almogren, A.; Abbas, A.; Khattak, H.A.; Din, I.U.; Guizani, M.; Zuair, M. Spammer Detection and Fake User Identification on Social Networks. *IEEE Access* **2019**, *7*, 68140–68152. [CrossRef]
20. Hansen, D.; Shneiderman, B.; Smith, M.A. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*; Morgan Kaufmann: Burlington, MA, USA, 2010.
21. Pournaki, A.; Gaisbauer, F.; Banisch, S.; Olbrich, E. The twitter explorer: A framework for observing Twitter through interactive networks. *Digit. Soc. Res.* **2021**, *3*, 106–118. [CrossRef]
22. Karmakharm, T.; Aletras, N.; Bontcheva, K. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Hong Kong, China, 3–7 November 2019; pp. 115–120.
23. Bevensee, E.; Aliapoulios, M.; Dougherty, Q.; Baumgartner, J.; Mccoy, D.; Blackburn, J. SMAT: The social media analysis toolkit. In Proceedings of the 14th International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 8–11 June 2020; Volume 14.
24. Hui, P.M.; Yang, K.C.; Torres-Lugo, C.; Monroe, Z.; McCarty, M.; Serrette, B.D.; Pentchev, V.; Menczer, F. Botslayer: Real-time detection of bot amplification on twitter. *J. Open Source Softw.* **2019**, *4*, 1706. [CrossRef]
25. Liu, X.; Li, Q.; Nourbakhsh, A.; Fang, R.; Thomas, M.; Anderson, K.; Kociuba, R.; Vedder, M.; Pomerville, S.; Wudali, R.; et al. Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 207–216.
26. Ram, R.; Kong, Q.; Rizoiu, M.A. Birdspotter: A Tool for Analyzing and Labeling Twitter Users. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Virtual Event, 8–12 March 2021; pp. 918–921.
27. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
28. Bruls, M.; Huizing, K.; Van Wijk, J.J. Squarified treemaps. In *Data visualization 2000*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 33–42.
29. Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **2014**, *9*, e98679. [CrossRef] [PubMed]
30. Majeed, S.; Uzair, M.; Qamar, U.; Farooq, A. Social Network Analysis Visualization Tools: A Comparative Review. In Proceedings of the 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 5–7 November 2020; pp. 1–6.
31. Pavlopoulos, G.A.; Paez-Espino, D.; Kyrpides, N.C.; Iliopoulos, I. Empirical comparison of visualization tools for larger-scale network analysis. *Adv. Bioinform.* **2017**, *2017*, 1278932. [CrossRef] [PubMed]

future internet

MDPI

Article

Aesthetic Trends and Semantic Web Adoption of Media Outlets Identified through Automated Archival Data Extraction

Aristeidis Lamprogeorgos, Minas Pergantis *, Michail Panagopoulos and Andreas Giannakoulopoulos *

Department of Audio and Visual Arts, Ionian University, 7 Tsirigoti Square, 49100 Corfu, Greece; a18labr@ionio.gr (A.L.); mpanagop@ionio.gr (M.P.)
* Correspondence: a19perg6@ionio.gr (M.P.); agiannak@ionio.gr (A.G.)

Abstract: The last decade has been a time of great progress in the World Wide Web and this progress has manifested in multiple ways, including both the diffusion and expansion of Semantic Web technologies and the advancement of the aesthetics and usability of Web user interfaces. Online media outlets have often been popular Web destinations and so they are expected to be at the forefront of innovation, both in terms of the integration of new technologies and in terms of the evolution of their interfaces. In this study, various Web data extraction techniques were employed to collect current and archival data from news websites that are popular in Greece, in order to monitor and record their progress through time. This collected information, which took the form of a website's source code and an impression of their homepage in different time instances of the last decade, has been used to identify trends concerning Semantic Web integration, DOM structure complexity, number of graphics, color usage, and more. The identified trends were analyzed and discussed with the purpose of gaining a better understanding of the ever-changing presence of the media industry on the Web. The study concluded that the introduction of Semantic Web technologies in online media outlets was rapid and extensive and that website structural and visual complexity presented a steady and significant positive trend, accompanied by increased adherence to color harmony.

Keywords: media industry; online media; news websites; Web data extraction; Semantic Web; Web aesthetics; Web archives; World Wide Web

Citation: Lamprogeorgos, A.; Pergantis, M.; Panagopoulos, M.; Giannakoulopoulos, A. Aesthetic Trends and Semantic Web Adoption of Media Outlets Identified through Automated Archival Data Extraction. *Future Internet* **2022**, *14*, 204. https://doi.org/10.3390/fi14070204

Academic Editor: Michael Sheng

Received: 8 June 2022
Accepted: 29 June 2022
Published: 30 June 2022

1. Introduction

Journalism was one of the first fields to make the transition from the physical realm to the online digital space, starting with the appearance of the Wall Street Journal in Bulletin Board Systems of the 1980s [1]. As soon as the World Wide Web started becoming popular, newspapers also started being published online, with the Palo Alto Weekly being available on the Web as early as January 1994 [1]. In the beginning, the printed content was being identically reproduced on the Web, but after a short period, some publications started being produced specifically for the Web, thus dramatically changing the way media outlets produced and disseminated their content according to Karlsson and Holt [2]. By the year 1999, more than 20% of American online newspaper content were Web originals, as claimed by Deuze's research the same year [3]. Ever since then, online media outlets have been capitalizing on the Web's power to provide journalistic content with traits that only it can offer, namely interactivity, immediacy, hypertextuality, and multimodality [2].

At the turn of the millennium, Tim Berners-Lee proposed the Semantic Web, an expansion of the World Wide Web that included content that can be retrieved and comprehended by machines, introducing the idea of a machine-readable Web [4]. In the field of Journalism, as Fernandez et al. point out, in order to cover the customers' needs for information freshness and relevance, the use of metadata became prevalent [5]. Moreover, the use of additional Semantic Web technologies as proposed by Fernandez et al. was set to increase both productivity and media outlet revenues [5]. Heravi and McGinnis proposed the use

of Semantic Web technologies, in tandem with Social Media technologies, to produce a new Social Semantic Journalism framework that combined technologies that could collaborate with each other in order to identify newsworthy user-generated journalistic content [6].

However, the evolution of the Web is not limited to content diffusion and machine-readability fields but also applies to the realms of aesthetics and usability. As Wu and Han point out, both aesthetics and usability display a strong relationship with the satisfaction of potential users [7]. King et al. [8] make the claim that a significant relationship exists between the visual complexity of a website and its influence on user first impressions. This is especially important with regard to media outlets since King's research specifically links increased visual complexity with the user's perception of informativeness and engagement cues [8]. This perceived informativeness is an important quality when associated with a news website. Besides complexity, usability and compatibility with multiple devices have also evolved through the progression of website layout techniques over the course of time as studied by Stoeva [9]. The way information is presented on a Web page is under constant change.

In addition to complexity and layout, color also plays an important role in influencing user impressions. On many occasions, researchers have established that the colors used on a website can elicit emotional reactions and feelings that can lead to outcomes concerning a website's perceived trustworthiness and appeal or even a visitor's overall satisfaction [7,10–12]. Talei proposes that these emotional responses are a result of human natural reactions to colors as encountered in natural life [12]. In addition to colors as individual factors eliciting an emotional reaction from users, White proposes that color schemes can also have a similar effect and proceeds to study the case of schemes using complementary colors [13] leading to conclusions about how specific complementary colors lead to increase in user pleasure.

In order to monitor how websites of media outlets evolve alongside the evolution of Web technologies and aesthetics, taking a look at contemporary websites only is not enough. Instead, what is needed is a comprehensive overview of each website's journey throughout the past decades. Brügger coined the term "website history" as a combination between media history and Internet history, where the individual website is considered the object of historical analysis instead of the medium [14]. The website then, playing the part of a historical document, is to be archived and preserved, and subsequently delivered as historical material [14]. This type of historical material is the means through which the aesthetic trends and Semantic Web adaptation of media outlets may be identified through means of archival data extraction.

The study presented in this article attempts to answer the following research questions:

RQ1. How has the integration of Semantic Web technologies (SWT) progressed in the last decades? When and to what extent were various technologies implemented?

RQ2. What are the trends in website aesthetics that can be identified concerning the complexity of Web pages, the usage of graphics, and the usage of fluid or responsive designs?

RQ3. What basic colors and coloring schemes are prevalent in website homepages? Did they change over the years and are there consistent trends that can be inferred by such changes?

In order to investigate these questions, large amounts of quantitative data were collected from actual public media outlets on the World Wide Web, based on their popularity in Greece. The past versions of these websites were retrieved through the use of a Web service offering archival information on websites. With that data in hand, a comprehensive understanding of the landscape of SWT adoption and general aesthetic trends can be attained. The method of collecting and analyzing that information will be presented in the following section.

2. Methodology

The research presented was conducted in four stages:

Stage 1: Media outlet websites were identified and selected based on their popularity in Greece.

Stage 2: Current and archival information from these websites was collected through the use of a website archive service. This information included the HyperText Markup Language (HTML) code of a website's homepage as well as a screenshot of that homepage.

Stage 3: Using a Web data extraction algorithm, information regarding the usage of SWTs, website complexity, graphic usage, and website repressiveness or fluidity was recorded.

Stage 4: Using an image analysis algorithm, information regarding the colors used was extracted from the websites' screenshots.

The methods and decision process behind each stage will be further detailed in this section. The quantitative data collected will be further presented in the results section.

2.1. Identifying Websites for Information Extraction

In order to reach safe conclusions regarding the evolution of media outlet websites through time, a large number of websites must be used, as well as multiple instances of each such website over the course of time. A large data set can lead to reliable results and create an impression that accurately represents reality. For that purpose, the archival Web service that was selected as the main provider of data concerning these websites was the Wayback Machine Internet Archive. As seen in the work of Gomes et al. [15], most Web archiving initiatives are national or regional. Out of the few international ones, the Internet Archive is both the largest and the oldest, dating back to 1996. It boasts over 625 million Web pages [16] which it provides to interested parties through its Wayback Machine. Using the Wayback Machine was considered the best way to collect a variety of instances for each studied website, which spanned over a representative period of time.

Another consideration, besides the number of instances, was which specific websites were to be targeted. A reliable metric of a media outlet's impact and visibility is its popularity based on digital traffic. Additionally, this popularity can ensure the existence of multiple instances of archived website data in Web archives. Based on that, a sample of the 1000 most popular websites was obtained from the SimilarWeb digital intelligence provider in the category of "News & Media Publishers" in Greece. SimilarWeb is a private company aiming to provide a comprehensive and detailed view of the digital world [17]. Information about a website's online market share, its global rank, and more were collected manually in the form of text files and using an algorithm scripted with PHP, this information was parsed and imported into a relational database powered by the MariaDB database management engine. This process is visually presented in Figure 1.

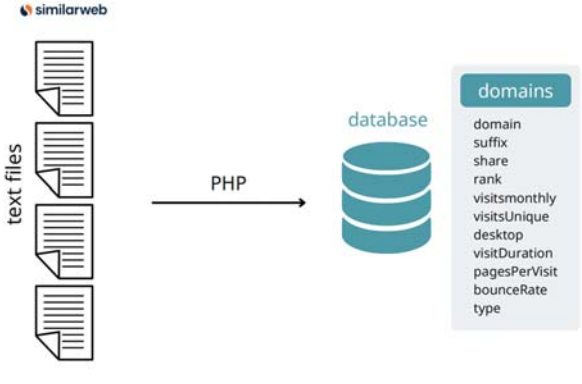


Figure 1. Visual representation of the process of website information collection.

Both international websites with a popular presence in Greece and popular Greek media outlets were included in the final list of websites to be investigated. Overall, the websites presented a varied mix including popular international online media outlets

(e.g., Yahoo, MSN, BBC, the NYTimes, etc.), popular Greek online media outlets such as (e.g., protothema.gr, iefimerida.gr, newsbomb.gr, etc.), a series of local news outlets with a popular online presence (e.g., typosthes.gr, thebest.gr, larissanet.gr, etc.), and more.

2.2. Collecting HTML Data and Screenshots of Each Relevant Website

Having established a good dataset of relevant websites, the next stage of this research was to collect HTML data and screenshots for each website for various different instances over the past few decades. An algorithm was developed in the PHP scripting language that inquired the Internet Archive's Wayback Machine for each of the websites collected in the previous stage, in order to obtain available instances for that specific website.

These inquiries were performed using the Wayback CDX (ChemDraw Exchange format) server Application Programming Interface (API). The CDX API is a tool that allows advanced queries that can be used to filter entries with high density instancing, in order to obtain instances for specific intervals. By using the API's ability to skip results in which a specific field is repeated, instance recovery was accomplished faster and more efficiently. For each instance of a website that is discovered in the Internet Archive's database, the API provides information on the domain name, the exact timestamp of the snapshot, the snapshot's year and month, the original Uniform Resource Locator (URL), the mime type of the data provided by the service and the current URL of the archived website on the Wayback Machine. This process of collecting instances is visually presented in Figure 2.

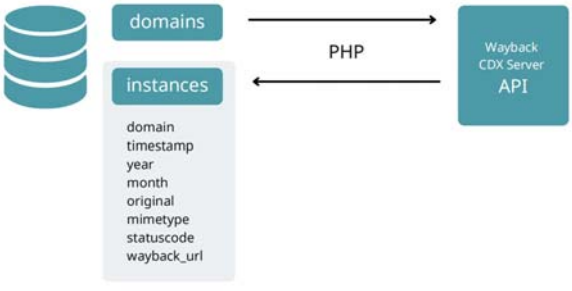


Figure 2. Visual representation of the process of instance information gathering.

Koehler, in their research, discovered that the half-life of a page is approximately 2 years [18]. Especially when it comes to structural or large-scale changes such as the ones that are being investigated in this research, it makes sense that they do not happen too often. With that in mind, for the purposes of this study, it was decided that one website instance per year was more than enough to record any significant changes. In order to accomplish this sampling, the timestamp field that was returned by the API was utilized. This field has 14 digits corresponding to the year, month, day, hour, minutes, and seconds that the instance was created. By instructing the API to exclude results that had the same first four digits in this field, the system returns exactly one snapshot per year as intended (if available). Out of a total of 1000 websites identified in stage one, 905 were discovered in the Internet Archive's databases and a grand total of 10,084 instances were discovered.

In order to acquire the HTML source code for each instance, an algorithm was developed in the PHP scripting language. This algorithm made use of the Wayback URL field that was collected during the instance information-gathering process to access the archival version of the website on the Wayback Machine. After accessing the instance, the algorithm proceeded to extract the source code and store it in an HTML file. The files were stored in a separate folder for each domain and their filenames represented the year and month of the instance. Before storing the source code into the HTML file, the application used string manipulation PHP functions to remove any part that belonged to the Wayback Machine's Web interface, in order to ensure that the end result was exclusively the original website's source code. This process is visually presented in Figure 3.

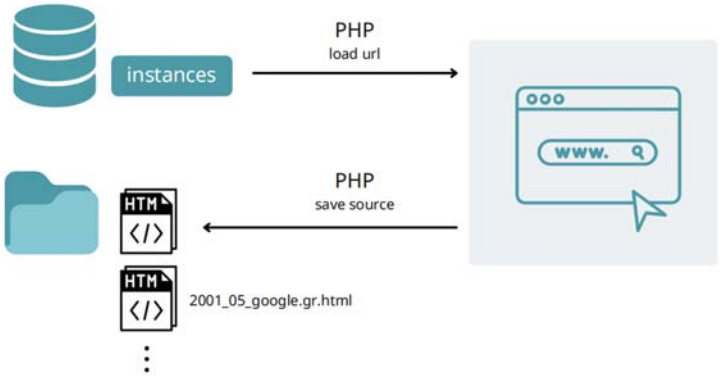


Figure 3. Visual representation of the HTML code collecting process.

The second important piece of information collected in this stage of our research besides the HTML source code is a screenshot of each website instance's homepage. The collected screenshots will be used to infer the color pallets of each instance and derive information from there. The plug-in UI.Vision RPA for the Chrome browser was used to acquire these screenshots. This plugin is a tool that allows the automation of various browser operations. The instructions for the automated process are provided to the plugin using JSON syntax. This enabled us to generate a vast series of instructions using an algorithm in the PHP scripting language. These instructions guide the plugin to open a website, pause for the time required for the website to load, capture a screenshot of the website, and then proceed to store the captured screenshot in a PNG image file with a filename indicating the year, month, and domain of the instance. This process is visually presented in Figure 4.

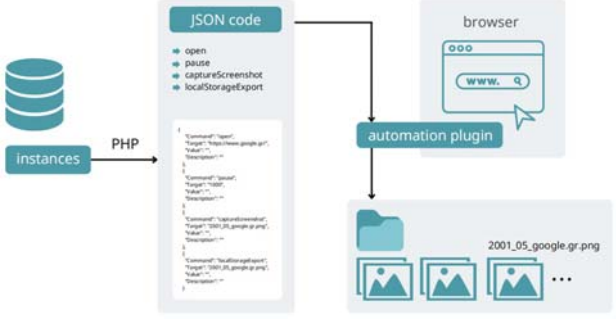


Figure 4. Visual representation of the screenshot collecting process.

This detailed overall process of gathering archival data related to website aesthetics can be extended for use in other fields and with different objectives that can be accomplished through knowledge of the HTML source code and a screenshot of a website instance and was presented in greater detail by Lamprogeorgos et al. in 2022 [19]. The complete process is visually presented in Figure 5. It should be noted that the process of collecting screenshots is much more resource and time intensive than the process of collecting HTML documents and for this reason, the analysis of screenshots was based on a random sample of 5402 website instance screenshots out of the 10084 total website instances. The screenshot sample was considered still large enough to lead to safe conclusions.

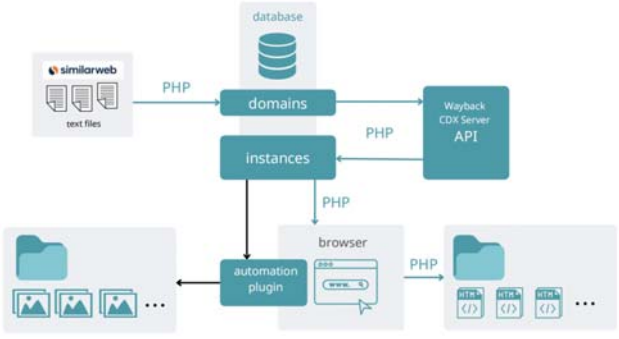


Figure 5. Visual representation of the complete instance data collecting process.

2.3. Collecting SWT and Aesthetics Data from the HTML Source Code

With the HTML files containing the source code of each website instance collected, the next step included the extraction of data from these files. The process of collecting this information was accomplished with the use of an algorithm developed in the PHP scripting language. This algorithm converted each HTML file into an entire HTML document through the use of PHP's DOMDocument Class. It then proceeded to collect information based on the various HTML elements and their attributes. This information was recorded into variables that can be divided into three categories: variables concerning Semantic Web technology adoption, variables concerning the homepage's complexity, and variables concerning the user interface's layout.

2.3.1. Semantic Web Technologies Adoption Variables

With the coming of HTML5 in 2008, a series of new structural elements were added [20] with the intention of not only providing structural insight, as normal HTML elements do but also contextual insight on what the content inside these elements represents. Fulanovic et al. indicate that usage of these elements is mainly intended for browsers and accessibility devices, and that it is up to the content creators to select the proper element to convey the contents of each part of their website [21]. These the elements are <*article*>, <*aside*>, <*details*>, <*figcaption*>, <*figure*>, <*footer*>, <*header*>, <*main*>, <*mark*>, <*nav*>, <*section*>, <*summary*>, and <*time*>. The data extraction algorithm traverses the Document Object Model (DOM) of each website and identifies the use of any of these elements and records it into the variable *html_var*.

The second variable concerning SWT adoption was *og* and it recorded whether a website made use of the Open Graph protocol to present itself in the form of a rich object. The protocol's intention is to make it possible for websites to be presented in a social graph and this is accomplished through a method compatible with W3Cs Resource Description Framework in attributes (RDFa) recommendation [22].

Another RDFa compatible system specifically designed for Twitter is called "Twitter Cards" [23] and whether it existed in a website instance was recorded in the *twitter* variable. Both the Open Graph and the Twitter Cards graphs create meta tag attributes that include information containing the Web page including a title, a short description, and a related image. Essentially, both Open Graph and Twitter Cards comprise Semantic Web applications that stem from the realm of Social Media as Infante-Moro et al. explain [24] and this connection they have with Social Media has influenced their popularity and their importance to websites' Semantic Web integration.

Although technically on the fence between Web 2.0 and Web 3.0, RDF Site Summary (RSS) feeds present one of the earliest attempts at Web syndication [25] and have hence been a long-time component of presenting Web pages and their content in a machine-readable manner. The variable *rss* records the existence of such feeds in a website instance.

Finally, the last SWT-related variable is *sch*, which records the existence of schema.org data structures in the website instance. The data structure schemas of the schema.org community, which is supported by various big names in Web technologies such as Microsoft, Google, Yahoo, and Yandex, aim to make it easier for developers to integrate sections of machine-readable information in their creations [26]. Their usage provides the flexibility of choosing between three formats: RDFa like Open Graph and Twitter Cards, Microdata, and JSON-LD.

Table 1 presents all SWT-related variables with a short description.

Table 1. Variables related to Semantic Web technologies.

Variable Name	Description
html_var	Records semantic HTML5 elements
og	Records the existence of an Open Graph RDFa graph
twitter	Records the existence of a Twitter Cards RDFa graph
rss	Records the existence of an RSS Feed
sch	Records the existence of a schema.org data structure

2.3.2. Aesthetics and Interface Variables

Visual complexity is a factor that plays an important role in the aesthetics of a website as discussed by Harper et al. [27], King et al. [8], and Chassy et al. [28]. Harper et al., in their work, supported that complexity as perceived by the users is influenced by structural complexity and presented a paradigm that related the complexity of an HTML document's DOM with how users subjectively judged complexity [27]. In a similar manner, the present study collected information regarding specific DOM elements, including both structural elements and graphical elements, in order to draw conclusions regarding the aesthetics of a website instance and how they evolved through time with regard to visual and structural complexity. In Figure 6, a screenshot of the homepage of popular European media outlet euronews.com (accessed on 1 January 2018), which displays a high amount of visual and structural complexity, is presented as an example.

In the *div_tags* variable the number of <*div*> elements was recorded while all hyperlinks were identified through the use of anchor elements <*a*> and recorded in the *a_tags* variable. Similarly, the various graphical components were measured using the *img_tags* variable to collect <*img*> elements, the *svg_tags* variable to collect scalable vector graphics elements (<*svg*>), the *map_tags* variable to collect image map elements (<*map*>), the *figure_tags* variable to collect figure semantic element (<*figure*>), the *picture_tags* variable to collect the art and responsive design oriented picture element (<*picture*>), and finally the *video_tags* variable to collect <*video*> elements.

The <*img*> tag is used to embed an image file in an HTML page. The image file can be of any Web-supported filetype such as compressed JPG files, animated GIF files, transparent PNG files, and even SVG files. An SVG element (<*svg*>) is a graphic saved in a two-dimensional vector graphic format that stores information that describes an image in text format based on XML. An image map consists of an image with clickable areas, where the user can click on the image and open the provided destination. The <map> tag can consist of more than one <*area*> element, which defines the coordinates and type of the area and any part of the image can be linked to other documents, without dividing the image. The <*figure*> tag is used to mark up a photo in the document on a Web page. Although the <*img*> tag is already available in HTML to display the pictures on Web pages, the <*figure*> tag is used to handle the group of diagrams, photos, code listing, etc. with some embedded content. The most common use of the <*picture*> element will be in responsive designs where instead of having one image that is scaled up or down based on the viewport width, multiple images can be designed to more nicely fill the browser viewport.

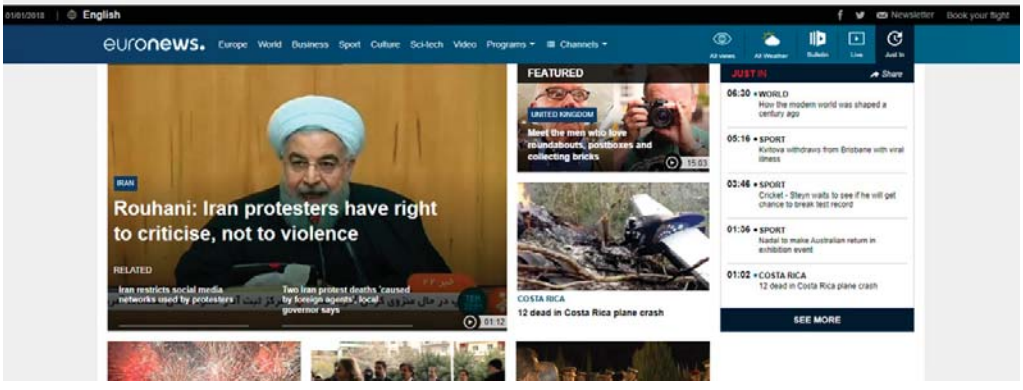


Figure 6. Example of the homepage of euronews.com displaying increased structural and visual complexity.

Besides visual complexity, modern website aesthetics and their interfaces are heavily influenced by the need to be presentable and easily usable on many different devices, operating at various different screen resolutions and aspect ratios. This has been achieved through the fluidity offered by using table elements to contain a website's structure and through the use of responsive design practices and frameworks. In order to study the trends in this area over time for each website instance, the number of table elements (<*table*>) was recorded in the *table_tags* variable. Additionally, the viewport meta element was investigated for each website instance as an indicator that the website is undertaking an effort towards supporting multiple screen resolutions and the results were recorded in the *mobile_scale* variable. Finally, two very popular responsive design frameworks were investigated. These were Bootstrap, an open source CSS framework developed by the Boostrap Team and operating under the Massachusetts Institute of Technology (MIT) license [29], and Foundation, a similar CSS framework also operating under the MIT license developed by ZURB [30]. In order to identify the frameworks, the algorithm tried to detect div elements with the grid "*row*" class and then proceeded to investigate for grid column elements through the various "*col-*" classes for Bootstrap and the "*columns*" and "*large-*" or "*small-*" classes for Foundation. Whenever the use of these frameworks was discovered, it was recorded in the *bootstrap* and *foundation* variables respectively.

Table 2 presents all visual complexity and layout structure-related variables with a short description.

Table 2. Variables related to visual complexity and layout structure.

Variable Name	Description
div_tags	Records the number of elements
a_tags	Records the number of <a> elements
img_tags	Records the number of <img> elements
svg_tags	Records the number of <svg> elements
map_tags	Records the number of <map> elements
figure_tags	Records the number of <figure> elements
picture_tags	Records the number of <picture> elements
video_tags	Records the number of <video> elements
table_tags	Records the number of <table> elements
mobile_scale	Records the number of elements
bootstrap	Records the existence of elements with classes used by the bootstrap framework
foundation	Records the existence of elements with classes used by the Foundation framework

2.4. Collecting Color Data from the Homepage Screenshot

Having amassed a large amount of website instance screenshots, we proceeded to use them in order to gain a better understanding of how news websites evolved through the last decades, in terms of empty space use and colors. Empty space (or white space, or negative space) is the unused space around the content and elements on a website, which designers used to balance the design of the website, organize the content and elements, and improve the visual experience for the user. Figure 7 presents an example of empty space from a homepage screenshot from the popular American media outlet nytimes.com (accessed 1 January 2014), where all the empty space has been marked with the use of the color orange.

Figure 7. Example of a homepage screenshot from nytimes.com with the empty space turned orange.

Figure 8 displays an example of the evolution of the homepage of the international media outlet hellomagazine.com throughout the last two decades. This collection of homepage screenshots exemplifies the visible evolution of structural and graphical complexity, as well as color and empty space usage, which comprise the metrics collected by our algorithms from each website instance, as detailed in Section 2.3.2 and in the current section.

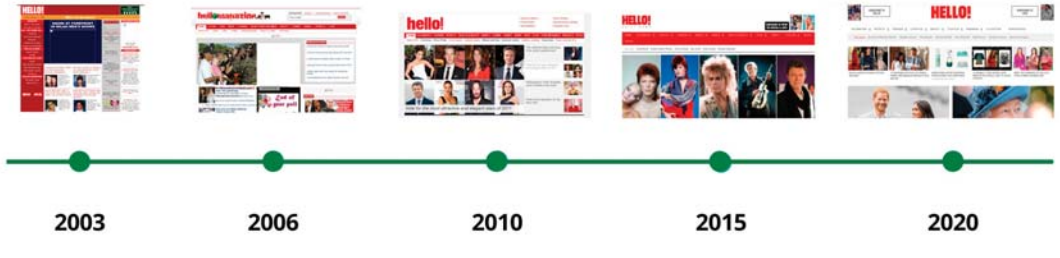


Figure 8. The evolution of the homepage of hellomagazine.com throughout the last two decades.

An algorithm was created that used the PHP scripting language and its native image handling capabilities to discover information regarding the use of color as presented by the screenshots. At first, the algorithm used image scaling and the *imagecolorat* function to identify and extract colors from a screenshot into the hexadecimal color code used by HTML5 and CSS3. Our work was based on the ImageSampler class developed by the Art of the Web [31]. All colors that took less than 3% of space on the screenshot are excluded from further analysis. In order to better study the remaining extracted colors, they were grouped based on their proximity to a primary, secondary, or tertiary color of the red yellow blue (RYB) color model.

As established by Gage in his work in the 1990s [32] the RYB color model incorporates subtractive color mixing and is one of the most popular color models, especially in design. By extension, it has become very useful in digital art and, of course, Web design since it can be used to identify colors that go well together. A major reason it was decided to convert the red green blue (RGB) based HTML hexadecimal colors to the RYB ones was to better study design schemes based on color relationships, as will be detailed below. The three primary colors of the RYB color wheel are red, yellow, and blue. Each combination of the three creates secondary colors which are orange, green, and purple. The tertiary colors are established through the combination of primary and secondary colors and they are red-orange, yellow-orange, yellow-green, blue-green, blue-purple, and red-purple. Additionally, black is achieved by combining all three primary colors and white through the lack of them.

The algorithm in this research used saturation to determine if a color is white: any color with less than 16% saturation was considered white. In a similar manner, brightness was used to identify black: any color with less than 16% brightness was considered black. Considering websites as a medium are presented across many different types of screens of various technologies, colors that are this close to black or white will most definitely be perceived as such by the average user. Additionally, the most used color on each website instance was considered to be the empty space color, meaning the color upon which all visual elements of the page appear.

In order to identify whether a color scheme (or color combination) is used in each website instance that uses colors besides black and white, an additional algorithm was developed in the PHP scripting language. This algorithm was designed to identify five major methods of color combination based on the RYB color wheel as presented in Figure 9:

- Monochromatic shades, tones, and tints of one base color
- Complementary colors that are on opposite sides of the color wheel
- Analogous colors that are side by side on the color wheel
- Triadic, three colors that are evenly spaced on the color wheel
- Tetradic, four colors that are evenly spaced on the color wheel

Figure 9. Major schemes of color combination based on the RYB color wheel.

The algorithm measured the minimum and maximum distance between the colors on the color wheel. Based on the number of colors and these two distances, conclusions can be drawn regarding the use of a harmonic color combination as presented in Figure 10.

Figure 10. Example of measuring distances on the RYB color wheel.

If the number of colors used is one, then the color scheme used is monochromatic. If the number of colors used is two, and if the maximum distance is lower than two, the analogous scheme is used, but if the maximum distance is greater than five, the complementary scheme is used. Similar conclusions can be drawn from the usage of three or four colors. If three colors are used and the minimum distance is greater than three, then the triadic color scheme is used. Similarly, if four colors are used and the minimum distance is greater than two, then the tetradic color scheme is implemented. The algorithm rejects any other situation and classifies it as a non-harmonic color combination.

Having obtained all relevant information through the steps described above, we proceeded to study the following:

- How many colors appear in the website instances on average by year besides black and white?
- How much each of the basic 14 colors of the RYB model is used in the website instances on average by year?
- How popular was the use of white, black, or colored empty space through the years?
- How popular were the different types of harmonic color combination schemes through the years?

The answers to these questions, alongside all other information collected throughout the stages of this research as presented in this section, are available in the results section below.

3. Results

The figures presented in this section focus on the various variables measured during the methodology section and how they shifted and changed in the last two decades from 2002 to 2022. Although some data since 1996 is available, it was deemed too small a sample to accurately present the information of that earlier era.

3.1. The Use of Semantic Web Technologies

Using the SWT adoption variable presented in Table 1 the usage of each technology was calculated as a percentage of the total website instances analyzed for each year. A graphic representation of the evolution of SWT adoption is presented in Figure 11. As we can see, the use of RSS feeds was the first introduction of machine-readability-related technologies in popular news media outlets circa 2004. Its use rose steadily in popularity until the early 2010s when the Open Graph technologies started rising to popularity alongside the use of HTML semantic elements. RSS usage stagnated around the same period forming a plateau in its curve. On the other hand, Open Graph and HTML semantic elements continued to rise in popularity until the late 2010s when they seem to have also plateaued but at significantly higher usage percentages. Twitter Cards and schema.org data structures start appearing en masse after 2014 and although their rate of diffusion is lowering they still display an upward trend.

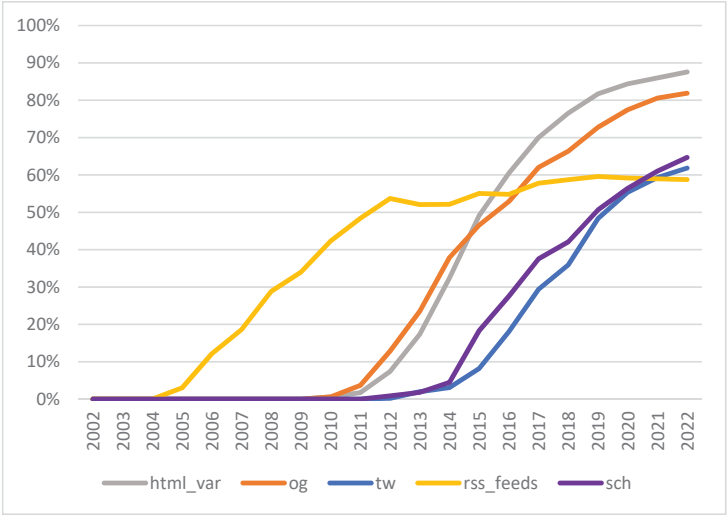


Figure 11. Percentage of websites using various Semantic Web technologies by year.

3.2. Website Complexity in Terms of Structural and Graphical Elements

Using the *div_tags* and *a_tags* variables as presented in Table 2, an overall impression of the evolution of structural complexity can be achieved. Figure 12 presents two curves, one for each variable, indicating what the average value for each variable was each year. The div element is more steep, beginning at a lower starting point in 2002 and ending above. Sometime after 2010, the average number of hyperlinks on news websites' front pages fell below the average number of div elements.

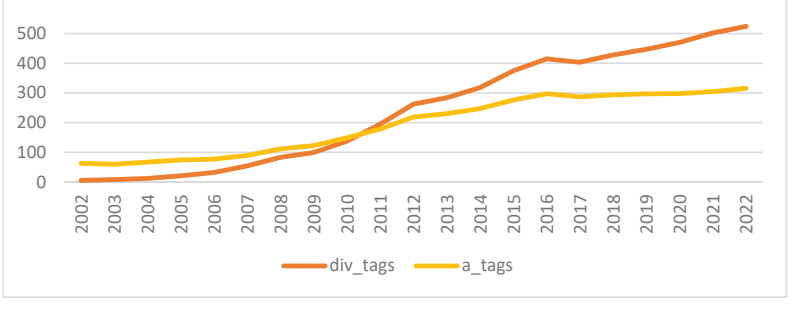


Figure 12. Website complexity as inferred through the average number of hyperlinks and div elements.

The total number of graphical elements identified every year as measured by the *img_tags*, *svg_tags*, *map_tags*, *figure_tags*, *picture_tags*, and *video_tags* of Table 2, were normalized by mapping them between the values of 0% and 100% (100% being the maximum detected number of elements) in order to make them independent of the number of website instances that were investigated, which were different for each year. Figure 13 presents the resulting curves that can be used to infer existing tendencies. Image elements in the form of an <*img*> tag display a slowly increasing trend, on par with the overall complexity increase depicted in Figure 12. From the early 2010s, a sharp rise appears in the figure tags which coincides with the rise of semantic elements usage as presented in Figure 11. The picture element also presents a similarly positive trend but lags 3–4 years behind since it is

more closely related to responsiveness, which will be studied further below. Image maps that were somewhat popular during the 2000s have been in fast decline and are all but extinct in modern websites. On the other hand, scalable vector graphics entered the field in the mid-2010s and their presence has been rising steeply ever since. Finally, the use of the video tag first appears after 2015 and has a peculiar double peak curve with its peaks in 2017 and 2022.

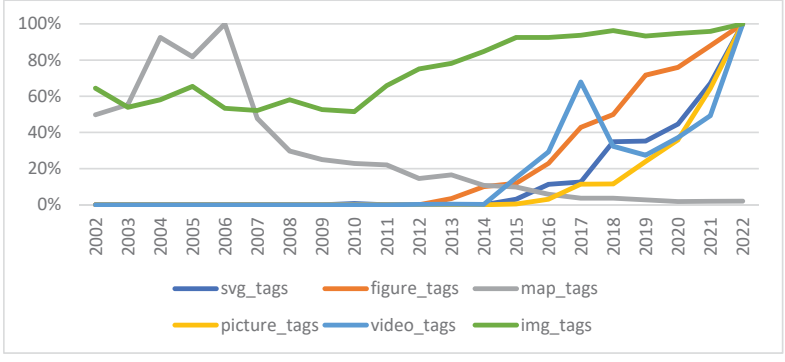


Figure 13. Graph presenting the number of graphical elements (normalized).

3.3. Fluidity and Responsiveness in Website Layouts

Fluid design as indicated by the *table_tags* variable measuring the existence of table elements, mobile device screen support as indicated by the *mobile_scale* variable, and responsive design as identified in the *bootstrap* and *foundation* variables are depicted in Figure 14. As seen clearly, fluid design after being very popular before 2008, has been in a steady decline since then, while, on the other hand, mobile support and responsive design have been rising rapidly in the 2010s.

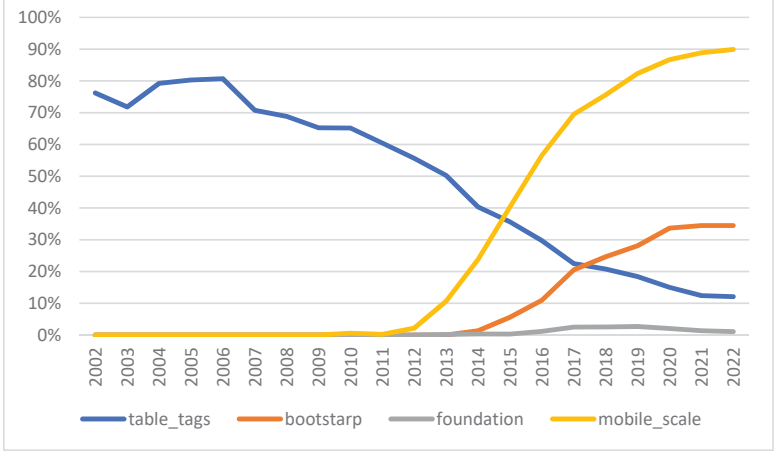


Figure 14. Graph presenting the usage of fluid or responsive design techniques.

3.4. Color and Empty Space Analysis Based on Website Homepage Screencaps

The number of different basic colors of the RYB color model that were used in each website instance provides a glimpse of the evolution of color-oriented aesthetical complexity. Figure 15 presents an area chart that depicts what percentage of each year's instances used

multiple colors and to what extent. Zero indicates no colors used besides black or white or other shades of the two, while each other number indicates how many basic colors were used besides black and white. It is made apparent that the use of fewer basic colors is more prevalent, although there is a trend as time goes by for the number of colors used to increase.

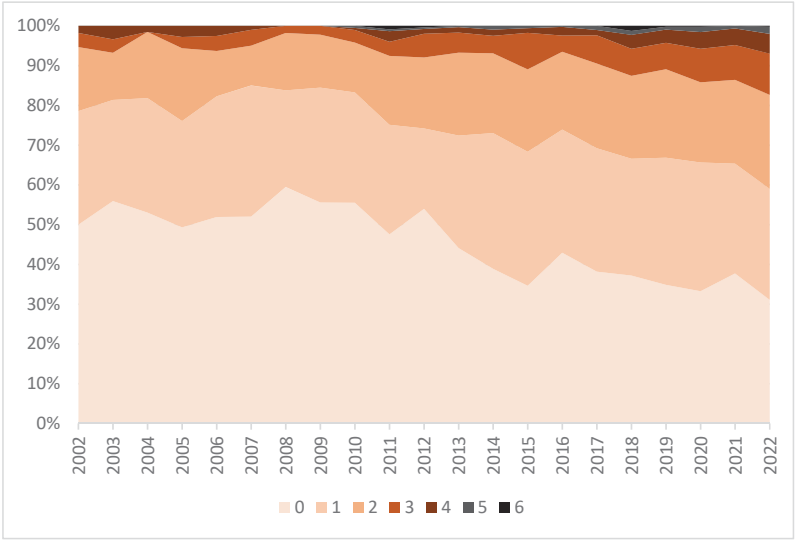


Figure 15. Number of basic RYB colors used besides black and white by year.

In order to form a more in-depth idea of which colors are used most, a graph showing the color usage of all 14 different colors is presented in Figure 16. As expected white is the most popular color, although displaying a somewhat negative trend. On the other side, black seems to display a positive trend, while other colors are much less used.

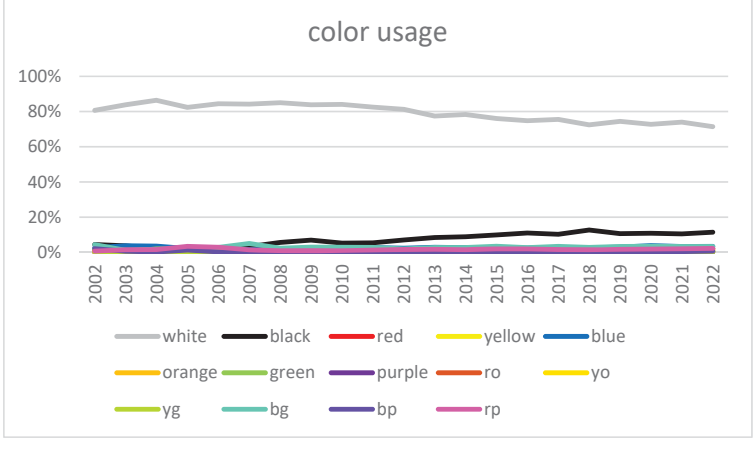


Figure 16. Basic RYB color usage by year.

Despite the fact that, as stated above, colors besides black and white are used much less, there might still be some trends to identify regarding the increase or decrease of their

use during these past decades. For this reason, their values were normalized between the values 0% and 100% where 100% was the value of the time when the color was most popular. These normalized values are presented in Figure 17.

(**a**) (**b**) (**c**)

Figure 17. Basic RYB color usage by year (normalized) for (**a**) black and white, (**b**) warm colors, and (**c**) cool colors.

In order to explore the usage of empty space in the collected website instances over the years, the most prevalent color detected in every instance was considered the empty space color. Figure 18 presents an area chart that depicts how the use of white and black as empty space colors evolved in the past decades. It is noticeable that while the use of white remains mostly steady, the use of black displays a small but significant positive trend.

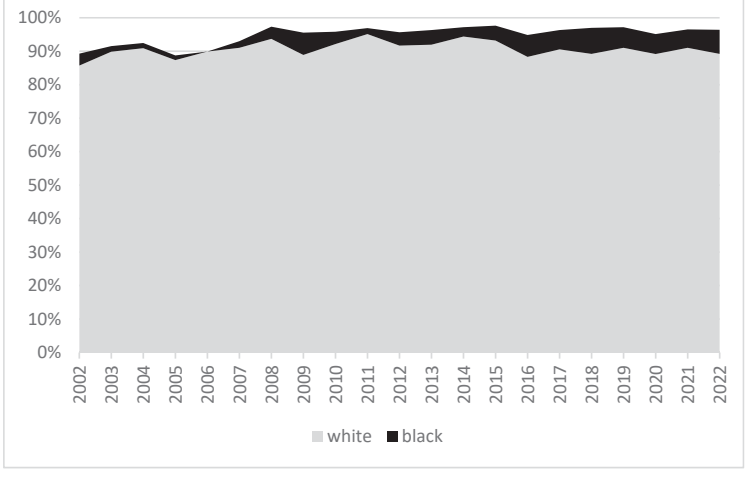


Figure 18. Usage of the colors white and black as empty space colors.

On some occasions, other basic RYB colors were detected as the empty space colors, though that percentage for each individual color was relatively small. Figure 19 presents

an area chart that depicts the use of other basic RYB colors as empty space colors over the past decades. An overall negative trend in the use of other colors is apparent, with a major drop in their usage around 2007 and 2008.

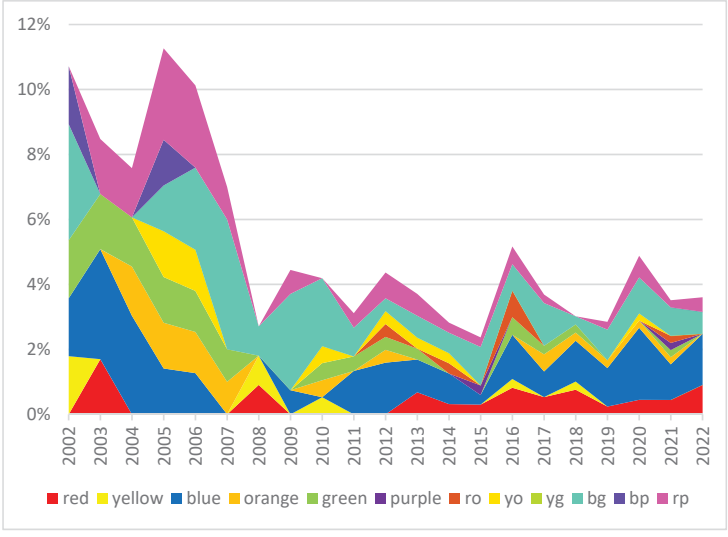


Figure 19. Usage of basic colors besides white and black as empty space colors.

Finally, as mentioned in the methodology section, the various color schemes that were identified in the collected website instances are presented in Figure 20. The monochromatic scheme, which uses one more basic color besides black and white, is prevalent throughout the studied period. At the same time there seem to be small positive trends in the usage of two complementary or two analogous colors. The use of three triadic colors is occasionally detectable but appears to be very limited throughout the past decades.

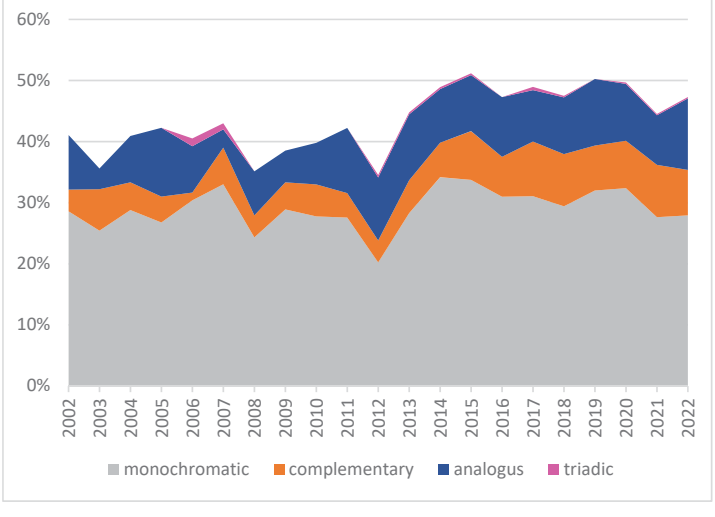


Figure 20. Usage of color schemes by year.

4. Discussion

4.1. RQ1: How Has the Integration of Semantic Web Technologies Progressed in the Last Decades? When and to What Extent Were Various Technologies Implemented?

As seen in Figure 11, the first signs of support for machine-readable content in the media outlets studied appeared in 2005. As Powers notes [25], the RSS2.0 specification was released in late 2002. From our findings, it appears that the practice of providing RSS2.0 feeds for public use started getting popular soon after. By 2010, almost half of the investigated website instances supported RSS.

Soon afterward, the adoption of Semantic HTML elements and Open Graph RDFa data begins. Back in the early 2010s, Fulanovic et al. [21] pointed out the importance of using the semantic elements instead of classes and ids to provide contextual information on websites and from our findings, it appears that this importance was acknowledged in the field of media outlets. In the same time period, the downward trajectory in traditional news media, alongside the rise of social media, as noted by Bajkiewicz et al. [33], dictated a shift from traditional media relations to a hybrid model, making most out of the Social Media environment. These facts support the extremely steep adoption curves that both HTML and OpenGraph displayed in Figure 11.

Within a few years, Semantic HTML elements gain presence in over 70% of website instances and by today almost 90% of website instances make use of at least some of them. In a similar style, more than 80% of modern websites are using Open Graph. Twitter Cards, which is also very closely connected to Social Media, follows a similar trend with Open Graph, just a couple of years later.

Schema.org data structures start getting identified as early as 2012 but did not start their steep climb in popularity until 2015. As noted by Muesel et al. [34], with schema.org's backing from major search engines, its adoption has been widespread. In a more recent study by Giannakoulopoulos et al., the usage of schema.org is found to be a competent predictor for Web traffic based on popularity in art and culture-related media outlets [35]. Based on all of the above it is apparent that SWT adaptation in media outlets is high and it is safe to assume that higher visibility and content diffusion are the main motivations behind this.

4.2. RQ2: What Are the Trends in Website Aesthetics That Can Be Identified Concerning the Complexity of Web pages, the Usage of Graphics, and the Usage of Fluid or Responsive Designs?

As clearly displayed in Figure 12, complexity, both as measured by <*div*> elements and as measured by hyperlinks are linearly increasing over the passage of time. The rate of increase is higher in <*div*> tags than it is in hyperlinks, but they are both overall pretty similar curves. As mentioned above, according to Harper et al. [27] HTML structural complexity is related to how visual complexity is perceived by the website visitors. According to King et al. [8], high levels of visual design complexity will result in both more favorable user first impressions and increase the users' perception of both visual informativeness and cues for engagement. This indicates a strong motivation for media outlet websites to present the user with such complexity. It should be noted that other studies such as Chassy et al. [28] and Harper et al. [27] had contradictory findings, with visual complexity appearing to negatively impact aesthetic pleasure. The difference may lie with the focus, which in one case was on informativeness and engagement and in the other cases on aesthetic pleasure. Users may judge a visually complex site as informative while a visually simpler site as beautiful. In the case of online media outlets, the first case is more in line with the website's intended purpose. In a similar manner as displayed in Figure 13, most graphical elements present positive trends throughout the years, which in turn lead to higher design complexity, which according to King et al. [8] can lead to much coveted favorable first impressions concerning informativeness. An exception is the image map (<*map*>) graphical element. The difference between this element and the rest is that it does not adjust to fluid or responsive layout design since its dimensions are fixed. With the rise of mobile Internet, it is expected that its usage has plummeted. The video element

also diverts from the norm, because there appears to be a fall in its usage after 2017 which has only recently been reversed. More detailed investigation through further research might shed some light on this matter, but it is noteworthy that in April 2018 was when Chrome and other Chromium-based browsers changed their autoplay policy to not allow video autoplay, in order to minimize the incentive for ad blockers and reduce mobile data consumption [36].

A major drawback of increased complexity, both in terms of structure and in terms of graphical elements, is that a large website file size negatively affects the website's loading times which can have an adverse effect on SEO [37]. However as time goes by, such technical limitations are overcome through the development of faster networks and devices with higher processing power which ensure fast loading times in increasingly large file sizes.

In terms of fluid and responsive design, Figure 14 paints a clear picture, with table elements diminishing while mobile support is increasing alongside grid-based responsive frameworks. The simple approach of table elements with fluid widths was a good first step into multiple screen resolution support, but with the mobile Internet becoming more prevalent, more than that was required. It is symbolically significant that in Figure 14, the table element curve crosses the mobile scale curve sometime in 2014, which was the year that mobile Internet usage exceeded that of the PC for the first time [38]. The *mobile_scale* variable reaches up to 90% in 2022, further reinforcing its significance. When it comes to specific frameworks, Bootstrap hovers above 30%, while Foundation is much lower. It is safe to assume that there are other ways to achieve repressiveness that our algorithm did not detect since there can be differences in framework keywords even from version to version of the same framework. Nevertheless, the rise in popularity of responsive web design tools is apparent from 2015 and onwards.

4.3. RQ3: What Basic Colors and Coloring Schemes Are Prevalent in Website Homepages? Did They Change over the Years and Are There Consistent Trends That Can Be Inferred by Such Changes?

As depicted in Figure 15, the number of RYB basic colors besides black and white used in website instances over the past decades is slowly but steadily increasing. Despite that, website instances only using black, white, and shades of gray were still the relative majority in 2022. Other than that, using one or two additional colors were also popular choices, and these three categories together constituted over 80% of our sample throughout the recent decades.

From a usage perspective, as seen in Figure 16, white and black are, of course, the most used colors. Usage of the other basic RYB colors is much less prevalent since, as we established in Figure 15, very few are usually used in each website instance. The normalized color usage graphs in Figure 17 can be used to identify trends in color usage. White seems to be displaying significant stability with a very limited decline noticed in the latest years. On the other hand, usage of black is increasing. The graphs for other colors do not display a clear pattern and can be quite erratic on a case-by-case basis. From the warm colors, red and red-orange seem to display a positive trend, while from the cool colors purple displays a similar pattern. Previous research by Alberts and Van der Geest has tried to link specific color usage on the Web with trustworthiness, finding blue, green, and red to be most positively linked to user perception of trust [10]. Bonnardel et al. investigated the various colors that appeal to users and Web designers and concluded that blue and orange were considered the most appealing [11]. Both these studies took place in 2011 after which both blue and orange seem to present a positive trend in our findings too. On the other hand, the use of green, which was considered second most related to trustworthiness by Alberts [10], has been steadily declining in our findings and is, in terms of usage, the third least used color overall. The blue-green tertiary color though has been found to be one of the most popular colors just behind white, black, and blue. Overall, the trends regarding specific color usage presented in Figure 17 can be used to draw rather limited conclusions. As Swasty et al. [39] note, user responses to color vary based on different

demographic factors such as age and gender. Additionally, what is important is not the specific color used but successfully utilizing that color to build brand identity. Despite that, Talaei describes that emotional response to specific colors is part of human nature [12] and our findings confirm that there are indications that the use of specific colors is a conscious design choice, aiming to create appeal. Nevertheless, further research work is required to draw safer conclusions.

On the matter of empty space color, white is the most popular choice, with black being a very distant second choice as seen in Figure 18. As Eisfeld and Kristallovich [40] present in a recent study, the light-on-dark color scheme, which is based on black as an empty space color, has been increasingly popular and has ushered the coming of "Dark Mode" in applications and websites. The intent of such an approach is reduced eye strain, as overall screen time for individuals increases [40]. On the other hand, as seen in Figure 19, using colors besides black and white as empty space colors has declined. A major factor in this development might be the fact that modern human–computer interaction design principles request a standard minimum contrast ratio which should be extended as discussed by Ahamed et al. [41] to improve both luminance and clarity. When using a color with above 16% brightness or saturation (which were the limits of black and white in our study) this contrast ratio is rather harder to achieve. Hence the media outlets' motivation to increase accessibility and usability might lead to the abandonment of using other basic colors as empty space colors.

Finally, Figure 20 presents the usage of the various color schemes. The general effort to produce visual complexity which can lead to improved first impressions from users [8], while at the same time maintaining color harmony, leads to the increase of complementary and analogous color schemes in the last decade. White [13], especially mentions that the use of complementary color schemes that evoke pleasure, can invoke positive attitudes towards advertisements and drive purchases. On the other hand, the triadic scheme still amounts to a very small portion of the website instances that were investigated in this research.

5. Conclusions

In this research, an innovative method was used to collect information from the HTML source code and homepage screenshots of a large number of websites, over a period of two decades, using data extraction techniques on archival data. The websites investigated were the top 1000 online media outlets based on Web traffic in Greece and included websites of both international media outlets and Greek national and local media outlets. The main goal of the study was to observe the course of these websites throughout the past decades, in regards to the adaptation of popular Semantic Web technologies and the aesthetic evolution of their interfaces, which included aspects concerning DOM structure and visual complexity, fluid, and responsive layout design techniques, and color usage and schemes.

The introduction of SWT in the websites was fast and extensive, with the main motivation behind it being the greater diffusion of media content. Structural and visual complexity displayed a steady but significant positive trend, aiming to achieve better first impressions while still maintaining performance across a plethora of devices. The rise of the mobile Internet guided the investigated websites to the adoption of responsive web design principles. An increase of visual complexity was also noted in the usage of colors, accompanied not only by an effort to better abide by the principles of accessibility, as established by the use of black as an empty space color but also by an effort to more closely adhere to color harmony through the use of color combinations.

The study's sample is large but does present limitations, in the sense that the criteria for selection were popularity on the Greek Web. Focusing on websites popular in a different country might have presented different results due to cultural or other factors. That being said, many of the studied websites were international media outlets, which would be popular in most of the world. An additional limitation of the research can be found in its focus on websites with high traffic, which might be inclined to adopt current technologies and trends more rapidly. Finding a more varied sample of media outlets that would

include low traffic or niche outlets could provide an interesting contrast. In the future, this research can be expanded to different fields of online activity, beyond news and media, and attempt to find comparable results. Additionally, focusing on regions with a large cultural distance to Greece could lead to conclusions regarding the connection between cultural identity and aesthetic trends. Moving forward, we will focus our future work on collecting information regarding a vast array of websites from different fields, beyond news outlets, while simultaneously adapting our metrics to better identify regional aesthetic trends, in order to contrast their development to global trends.

The World Wide Web is a constantly evolving entity that is influenced both by the rise and fall of technologies and by the continuous evolution of human nature through cultural trends, global events, and globalization in general. Studies of the Web's past and its course through time can provide valuable knowledge, pertaining not only to the present but hopefully preparing us for the future. The advancements of the Semantic Web and the aesthetic evolution of user interfaces can be useful tools at the disposal of every online media outlet, both established and new, and can lead to the overall betterment of the undeniable services they provide.

Author Contributions: Conceptualization, A.L. and A.G.; data curation, A.L. and M.P. (Minas Pergantis); formal analysis, A.L., M.P. (Minas Pergantis), and M.P. (Michail Panagopoulos); investigation, A.L. and M.P. (Minas Pergantis); methodology, A.L., M.P. (Minas Pergantis), M.P. (Michail Panagopoulos), and A.G.; project administration, A.G.; resources, M.P. (Minas Pergantis) and A.G.; software, A.L. and M.P. (Minas Pergantis); supervision, M.P. (Michail Panagopoulos) and A.G.; validation, M.P. (Michail Panagopoulos) and A.G.; visualization, A.L.; writing—original draft, A.L. and M.P. (Minas Pergantis); writing—review and editing, A.L., M.P. (Minas Pergantis), and M.P. (Michail Panagopoulos). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Zenodo at (https://doi.org/10.5281/zenodo.6624915, accessed on 7 June 2022), reference number (10.5281/zenodo.6624915).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Díaz Noci, J. A history of journalism on the Internet: A state of the art and some methodological trends. *RIHC Rev. Int. Hist. Comun.* **2013**, *1*, 253–272. [CrossRef]
2. Karlsson, M.; Holt, K. Journalism on the Web. In *Oxford Research Encyclopedia of Communication*; Oxford University Press: Oxford, UK, 2016.
3. Deuze, M. Journalism and the Web: An analysis of skills and standards in an online environment. *Gazette* **1999**, *61*, 373–390. [CrossRef]
4. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
5. Fernandez, N.; Blazquez, J.M.; Fisteus, J.A.; Sanchez, L.; Sintek, M.; Bernardi, A.; Fuentes, M.; Marrara, A.; Ben-Asher, Z. News: Bringing semantic web technologies into news agencies. In Proceedings of the International Semantic Web Conference, Athens, GA, USA, 5–9 September 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 778–791.
6. Heravi, B.R.; McGinnis, J. Introducing social semantic journalism. *J. Media Innov.* **2015**, *2*, 131–140. [CrossRef]
7. Wu, O.; Han, M. Screenshot-based color compatibility assessment and transfer for Web pages. *Multimed. Tools Appl.* **2018**, *77*, 6671–6698. [CrossRef]
8. King, A.J.; Lazard, A.J.; White, S.R. The influence of visual complexity on initial user impressions: Testing the persuasive model of web design. *Behav. Inf. Technol.* **2020**, *39*, 497–510. [CrossRef]
9. Stoeva, M. Evolution of Website Layout. In Proceedings of the Techniques Anniversary International Scientific Conference "Computer Technologies and Applications", Pamporovo, Bulgaria, 15–17 September 2021.
10. Alberts, W.; Van der Geest, T. Color Matters: Color as Trustworthiness Cue in Web Sites. *Tech. Commun.* **2011**, *58*, 149–160.
11. Bonnardel, N.; Piolat, A.; Le Bigot, L. The impact of colour on Website appeal and users' cognitive processes. *Displays* **2011**, *32*, 69–80. [CrossRef]
12. Talaei, M. Study of human reactions than color and its effects on advertising. *Int. J. Account. Res.* **2013**, *42*, 1–9. [CrossRef]

13. White, A.E.R. Complementary Colors and Consumer Behavior: Emotional Affect, Attitude, and Purchase Intention in the Context of Web Banner Advertisements. Ph.D. Thesis, Universidade Nova de Lisboa, Caparica, Portugal, 2018. Available online: http://hdl.handle.net/10362/52273 (accessed on 7 June 2022).
14. Brügger, N. The archived website and website philology. *Nord. Rev.* **2008**, *29*, 155–175. [CrossRef]
15. Gomes, D.; Miranda, J.; Costa, M. A survey on web archiving initiatives. In Proceedings of the International Conference on Theory and Practice of Digital Libraries, Berlin, Germany, 25–29 September 2011; Springer: Berlin/Heidelberg, Germany.
16. Internet Archive. About the Internet Archive. Available online: https://archive.org/about/ (accessed on 1 June 2022).
17. SimilarWeb. We Are the Official Measure of the Digital World. Available online: https://www.similarweb.com/corp/about/ (accessed on 1 June 2022).
18. Koehler, W. Web page change and persistence—A four-year longitudinal study. *J. Am. Soc. Inf. Sci. Technol.* **2002**, *53*, 162–171. [CrossRef]
19. Lamprogeorgos, A.; Pergantis, M.; Giannakoulopoulos, A. A methodological guide to gathering archival data related to website aesthetics. In Proceedings of the 4th International Conference Digital Culture & AudioVisual Challenges, Corfu, Greek, 13–14 May 2022. *Under publication*.
20. HTML Semantic Elements. Available online: https://www.w3schools.com/html/html5_semantic_elements.asp (accessed on 1 June 2022).
21. Fulanovic, B.; Kucak, D.; Djambic, G. Structuring documents with new HTML5 semantic elements. In Proceedings of the 23rd DAAAM International Symposium on Intelligent Manufacturing and Automation, Zadar, Croatia, 24–27 October 2012; Volume 2, pp. 723–726.
22. The Open Graph Protocol. Available online: https://ogp.me/ (accessed on 1 June 2022).
23. About Twitter Cards. Available online: https://developer.twitter.com/en/docs/twitter-for-websites/cards/overview/abouts-cards (accessed on 1 June 2022).
24. Infante-Moro, A.; Zavate, A.; Infante-Moro, J.C. The influence/impact of Semantic Web technologies on Social Media. *Int. J. Inf. Syst. Softw. Eng. Big Co.* **2015**, *2*, 18–30.
25. Powers, S. *Practical RDF*; O'Reilly Media, Inc.: Cambridge, MA, USA, 2003; pp. 10, 254.
26. Mika, P. On schema.org and why it matters for the web. *IEEE Internet Comput.* **2015**, *19*, 52–55. [CrossRef]
27. Harper, S.; Jay, C.; Michailidou, E.; Quan, H. Analysing the visual complexity of web pages using document structure. *Behav. Inf. Technol.* **2013**, *32*, 491–502. [CrossRef]
28. Chassy, P.; Fitzpatrick, J.V.; Jones, A.J.; Pennington, G. Complexity and aesthetic pleasure in websites: An eye tracking study. *J. Interact. Sci.* **2017**, *5*, 3. [CrossRef]
29. Bootstrap Team. Bootstrap—The Most Popular HTML, CSS and JS Library in the World. Available online: https://getbootstrap.com/ (accessed on 1 June 2022).
30. ZURB. Foundation—The Most Advanced Responsive Front-End Framework in the World. Available online: https://get.foundation/ (accessed on 1 June 2022).
31. Art of the Web. PHP: Extracting Colours from an Image. Available online: https://www.the-art-of-web.com/php/extract-image-color/ (accessed on 1 June 2022).
32. Gage, J. *Color and Meaning: Art, Science, and Symbolism*; University of California Press: Oakland, CA, USA, 1999.
33. Bajkiewicz, T.E.; Kraus, J.J.; Hong, S.Y. The impact of newsroom changes and the rise of social media on the practice of media relations. *Public Relat. Rev.* **2011**, *37*, 329–331. [CrossRef]
34. Meusel, R.; Bizer, C.; Paulheim, H. A web-scale study of the adoption and evolution of the schema. org vocabulary over time. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, New York, NY, USA, 13–15 July 2015; pp. 1–11.
35. Giannakoulopoulos, A.; Pergantis, M.; Konstantinou, N.; Kouretsis, A.; Lamprogeorgos, A.; Varlamis, I. Estimation on the Importance of Semantic Web Integration for Art and Culture Related Online Media Outlets. *Future Internet* **2022**, *14*, 36. [CrossRef]
36. Beaufort, F. Autoplay Policy in Chrome. 2017. Available online: https://developer.chrome.com/blog/autoplay/ (accessed on 1 June 2022).
37. Chotikitpat, K.; Nilsook, P.; Sodsee, S. Techniques for improving website rankings with search engine optimization (SEO). *Adv. Sci. Lett.* **2015**, *21*, 3219–3224. [CrossRef]
38. Murtagh, R. Mobile Now Exceeds PC: The Biggest Shift Since the Internet Began. *Search Engine Watch* **2014**. Available online: https://www.searchenginewatch.com/2014/07/08/mobile-now-exceeds-pc-the-biggest-shift-since-the-Internet-began/ (accessed on 1 June 2022).
39. Swasty, W.; Adriyanto, A.R. Does color matter on web user interface design. *CommIT (Commun. Inf. Technol.) J.* **2017**, *11*, 17–24. [CrossRef]
40. Eisfeld, H.; Kristallovich, F. The rise of dark mode: A qualitative study of an emerging user interface design trend. *Jönköping* **2020**. Available online: http://hj.diva-portal.org/smash/get/diva2:1464394/FULLTEXT01.pdf (accessed on 1 June 2022).
41. Ahamed, M.; Bakar, Z.; Yafooz, W. The Impact of Web Contents Color Contrast on Human Psychology in the Lens of HCI. *Int. J. Inf. Technol. Comput. Sci.* **2019**, *11*, 27–33. [CrossRef]

future internet

MDPI

Article

Modeling and Validating a News Recommender Algorithm in a Mainstream Medium-Sized News Organization: An Experimental Approach

Paschalia (Lia) Spyridou *, Constantinos Djouvas and Dimitra Milioni

Department of Communication and Internet Studies, Cyprus University of Technology, Saripolou 33, 3036 Limassol, Cyprus
* Correspondence: l.spyridou@cut.ac.cy

Abstract: News recommending systems (NRSs) are algorithmic tools that filter incoming streams of information according to the users' preferences or point them to additional items of interest. In today's high-choice media environment, attention shifts easily between platforms and news sites and is greatly affected by algorithmic technologies; news personalization is increasingly used by news media to woo and retain users' attention and loyalty. The present study examines the implementation of a news recommender algorithm in a leading news media organization on the basis of observation of the recommender system's outputs. Drawing on an experimental design employing the 'algorithmic audit' method, and more specifically the 'collaborative audit' which entails utilizing users as testers of algorithmic systems, we analyze the composition of the personalized MyNews area in terms of accuracy and user engagement. Premised on the idea of algorithms being black boxes, the study has a two-fold aim: first, to identify the implicated design parameters enlightening the underlying functionality of the algorithm, and second, to evaluate in practice the NRS through the deployed experimentation. Results suggest that although the recommender algorithm manages to discriminate between different users on the basis of their past behavior, overall, it underperforms. We find that this is related to flawed design decisions rather than technical deficiencies. The study offers insights to guide the improvement of NRSs' design that both considers the production capabilities of the news organization and supports business goals, user demands and journalism's civic values.

Keywords: news personalization; news recommender systems; algorithmic design; algorithmic journalism; algorithmic agenda; beyond accuracy

Citation: Spyridou, P.(.; Djouvas, C.; Milioni, D. Modeling and Validating a News Recommender Algorithm in a Mainstream Medium-Sized News Organization: An Experimental Approach. *Future Internet* **2022**, *14*, 284. https://doi.org/10.3390/fi14100284

Academic Editors: Luis Javier Garcia Villalba, Andreas Veglis and Charalampos Dimoulas

Received: 8 August 2022
Accepted: 26 September 2022
Published: 29 September 2022

1. Introduction

More than eight in ten Americans consume news from digital devices, with 60% claiming to do so often [1]. Similar trends are documented in Europe, yet recent research [2] found that the youngest cohort represents a more casual, less loyal news user. Social natives tend to heavily rely on social media for news, while both digital and social natives share a weak connection with brands, making it harder for media organizations to attract and engage them. At the same time, younger audiences are also particularly suspicious and less trusting of all information provided by news outlets (p. 45). These findings reflect the ongoing turmoil of the news media industry. In a high-choice media environment where multiple players offer news and users can access news content from a variety of pathways, and in many different modalities [3], media organizations find it increasingly hard to woo and retain users' attention. Myllylahti (2020) [4] defines attention as "a scarce and fluid commodity which carries monetary value" (p. 568); it is based on individual user interaction which can be measured and analyzed through detailed web metrics and analytics and exchanged for revenue [5,6]. As a result, news media resort to multiple strategies and techniques aiming to control what news people pay attention to and the conditions under which they do so in order to generate revenue [7]. A prominent technique increasingly used

by news organizations is News Recommending Systems (NRSs). NRSs are algorithmic tools that filter incoming streams of information according to the users' preferences or point them to additional items of interest. Harnessing the deluge of big data and machine learning [8], these technical systems aggregate, filter, select and prioritize information [9]. The *Daily Me* project foreseen by Nicholas Negroponte [10] in 1995 is becoming common practice as news publishers are increasingly experimenting with recommender systems to extent the provision of personalized news [11] hoping to increase their sites' 'stickiness', capture user data and reduce their dependence on external suppliers of such information [12].

However, their job is not easy; in today's high-choice media environment, attention shifts easily between platforms and (news) sites [4] and is greatly affected by algorithmic technologies [9,13]. More importantly, the task of recommending appropriate and relevant news stories to readers proves particularly challenging; in addition to the technical and design challenges associated with the news recommendations of most media offerings, personalized systems of news delivery present a special case given the impact of news for an informed citizenry [14,15].

Taking into consideration the business challenges of media organizations and the particularities of news content, the present study probes the algorithmic design of a news recommender introduced in the newsroom of a leading news portal in Cyprus. The study follows a design-oriented approach [16] aiming to identify the implicated parameters enlightening the underlying functionality of the algorithm and evaluate the NRS in hand to offer insights that can guide the improvement of NRS that support and align with business goals, user demands and journalism's civic values.

2. Problem Definition and Motivation: Modeling a News Recommending System

The collapse of the traditional advertising model [17], the web's free news culture [18] and the growing role of the platforms in news distribution [19,20] have made it increasingly difficult for news organizations to cope with editorial and commercial standards [21]. Initially, publishers saw the platforms as an ally to help them boost content visibility and brand awareness; soon, they realized that this evolving publisher–platform partnership is unequal; platforms wield more power over user data and earn significantly more advertising revenue than publishers [22,23]. Amid increasing demands to woo and retain users' attention, the use of algorithmic and data-driven techniques is gaining more and more ground in the media industry; they are used to automate workflows by modeling human-centered practices, thereby assigning new roles to both machines and media professionals [24,25].

News recommendations bear substantial benefits for all parties involved: first, they comprise an effective content monetization tool in terms of building traffic, engagement and loyalty [26]. Second, they help readers discover the depth of the outlets' reporting; The *New York Times*, for example, publishes approximately 250 stories per day. Algorithmic curation is used to propose content facilitating users to encounter news stories that might prove helpful and interesting to them and might otherwise not find while motivating the medium to keep producing a wide range of content [27], as the tailored delivery of news allows republishing content on a much broader scale. Finally, NRS have shown to be a useful tool for helping users deal with information overload [28]. Users are bombarded with news and information from different news outlets, social network posts, notifications, emails, etc., a situation which affects their attention span while making it increasingly difficult for them to find content of interest, at the right moment, in the right form [29].

Despite the acknowledged merits of news personalization, the technical challenge for offering effective recommendations is high [9,30] and particularly expensive [31]. Karimi and his colleagues [32] provide a comprehensive review of the many challenges associated with algorithmic accuracy and user profiling in news recommender systems. However, apart from accuracy and user profiling which comprise typical algorithmic challenges, news recommendation systems present a special case considering their civic function in the direction of an informed citizenry [33]. From a normative perspective, news provides the necessary information for citizens to think, discuss and decide wisely, to participate

in political life and thus democracy to function [34]. For that reason, much work on news recommendations focuses on exposure diversity as a design principle for news recommender systems (see [35–37]). The idea is that for a functioning democracy, users should encounter a variety of themes, opinions and ideas. Helberger and her colleagues [38] argue that "recommendation systems can be instrumental in realizing or obstructing public values and freedom of expression in a digital society; much depends on the design of these systems" (p. 3).

Recent diversity preoccupations echo older concerns arguing that the retreat from human editorial decision making in favor of machine choices might prove problematic. Pariser's [39] widely known hypothesis over filter bubbles refers to algorithmic filtering that tends to promote tailored content based on users' pre-existing attitudes, interests and prejudices, leading citizens into content 'bubbles'. In other words, personalized news offerings may result in people losing the common ground of news [40] by producing "different individual realities" [41] that hinder debate and amplify audience fragmentation and polarization [42]. Relevant work so far provides mixed results. While some studies provide evidence supporting the idea of filter bubble effects (e.g., amplification of selective exposure, negative effects on knowledge gain) [43,44], others argue that these fears are severely exaggerated [45,46]. Considering the social impact of news delivery, and Broussard's [30] argument that the mathematical logic of computers may do well in calculating but often falls short in complex tasks with social or ethical consequences, the possibility that news recommenders can lead to information inequalities or amplify existing biases and thus undermine the democratic functions of the media cannot be excluded [33]. Going a step further, Helberger and her colleagues [35] argue that filtering and recommendation systems can, at least in principle, be designed to offer a personalized news diet which both serves (assumed) individual needs and interests while catering for the provision of a democratic news diet. Recent work by Vrijenhoek and his colleagues [47] formulates the problem explicitly: the question is whether news recommenders are merely designed to generate clicks and short-term engagement or if they are programmed to balance relevance along with helping users discover diverse news and not miss out on important information (p. 173).

Figure 1 depicts the proposed model as a generic approach to describe NRS data flows and processes. The news organization produces a number of news items to be delivered. These items are classified into thematic categories or assigned multiple tags. At the same time, preferences set by the users and/or their previous browsing experience are correlated and matched with the features of the associated stories. Then, the algorithm ranks the matching stories, aiming to deliver accurate recommendations. When developing an effective NRS, one must also consider the beyond-accuracy aspects to evaluate the quality of news recommendations [14]. Therefore, apart from accuracy, the model incorporates the elements of diversity, serendipity and novelty (further defined in the following section) as input features used to limit the possibility of echo-chamber effects while enhancing both users' news experience and their engagement. Overall, the proposed model incorporates baseline data flows and processes that common NRS employ, elaborating on the civic role of journalism (reddish routes) through novel metrics to make the underlying mechanisms more robust and useful.

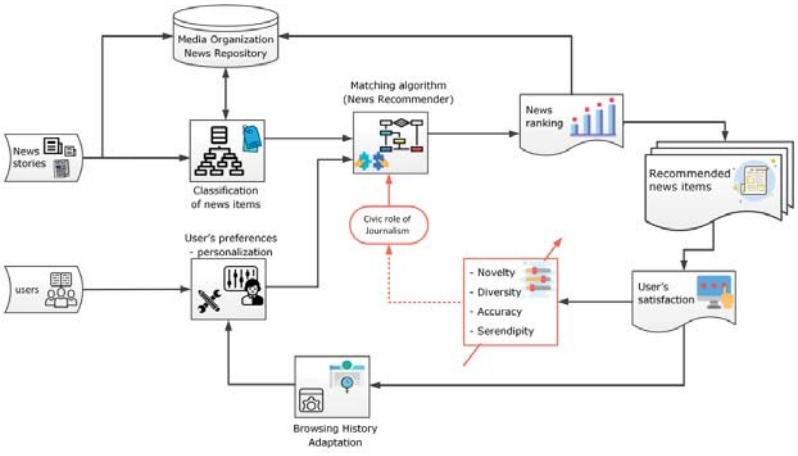


Figure 1. The proposed model incorporates baseline NRS data flows and processes, elaborating on the civic role of journalism (reddish routes) through novel metrics to make the underlying mechanisms more robust and useful.

3. Design Challenges for News Recommenders

An algorithm can be defined as a series of steps undertaken to solve a particular problem or accomplish a defined outcome [48]. In the context of news recommenders, the task of algorithms is to structure and order the pool of available news according to predetermined principles [49]. Algorithms though are embedded with values, biases or ideologies [50] that can influence an effective and unbiased provision of information [40]. Diakopoulos [48] speaks of algorithmic power premised in algorithms' atomic decisions, including prioritization, classification, association and filtering. Some algorithmic power may be exerted intentionally, while other aspects might be unintended side effects rooted in design decisions, objective descriptions, constraints and business rules embedded in the system, major changes that have happened over time, as well as implementation details that might be relevant (p. 404). News recommender systems are often classified into four main categories:

(1) *Collaborative filtering*: Items are recommended to a user based upon values assigned by other people with similar taste. Users are grouped into clusters on the basis of their preferences, habits or content ranking [51]; in practice, collaborative filtering automates the process of 'word-of-mouth' recommendations [52] and is found to be the most common approach in the recommender system literature [32].

(2) *Popularity filtering*: In this case, items are rated for their general popularity among all users; it is the simplest approach, as all users receive the same recommendations, potentially leading to 'popularity biases' and 'bandwagon effects' in which consumers gravitate toward already popular items [33].

(3) *Content-based filtering:* The main idea here is to create clusters of content and associate these clusters with user profiles [51]. Content-based filtering tries to recommend items similar to those a given user has liked in the past based on similarity scores of a user toward all the items [53].

(4) *Hybrid approaches:* Often, news recommender systems use a hybrid approach combining content-based filtering and collaborative filtering (Karimi et al., 2018), also including other methods such as weighing items by recency or pushing content that has specific features (e.g., paid content) [49].

The aforementioned types describe data-driven algorithms, but most news recommender systems employ additional rules which basically shape the overall design of the system. In most cases, these rules are jointly decided by the engineers and the editorial

team of the news organization [16]. So, while in rule-based systems, the automated process is carefully designed to reflect the choices and the criteria that are set by the creators; in data-driven systems, algorithmic bias is less direct, as it is caused by the attributes of the available data that are used to build the decision-making model through the algorithmic process.

When designing NRSs, several issues need to be taken under consideration. Most recommender algorithms are based on a topic-centered approach aiming to meet the different interests of users [54]. However, this approach can prove ineffective considering the specific nature of news items compared to other media offerings [55]. To begin with, news classification (tagging) is a difficult task, as news items may belong to more than one news category. Most often, news organizations follow their own typology of tagging and classifying news stories, which is a practice with a substantial degree of subjectivity. Additionally, most news items have a short lifespan, and thus, it is necessary to process them as fast as possible and start recommending them because their information value degrades [51]. In addition to the element of recency, the popularity of news items may differ dramatically, thereby rendering the traditional recommendation methods unsuccessful [32]. Figure 2 summarizes the main design challenges associated with the particularities of news as a media offering. Furthermore, news recommender systems must deal with a large and heterogenous corpus of news items generating scalability issues. A common strategy for solving scalability is clustering; effective clustering though requires a very thorough classification of news stories and detailed user profiling based on the reading behavior of users and their short-term and long-term profiles [54], which often proves a difficult task. The cold start problem stemming from insufficient user information to estimate user preferences is a common challenge in NRS [32,56]. User registration, among others, comprises a standard method to overcome it. Lavie and her colleagues [57], however, found significant differences between declared and actual interests in news topics, especially in broad news categories containing many subtopics (for instance, politics). They concluded that users cannot accurately assess their interest in news topics and argue that news recommender systems should apply different filtering mechanisms for different news categories. In other words, the depth of personalization should be adjusted to cater for both declared interests and assumed interests of important events (see Figure 3, Challenges associated with profiling).

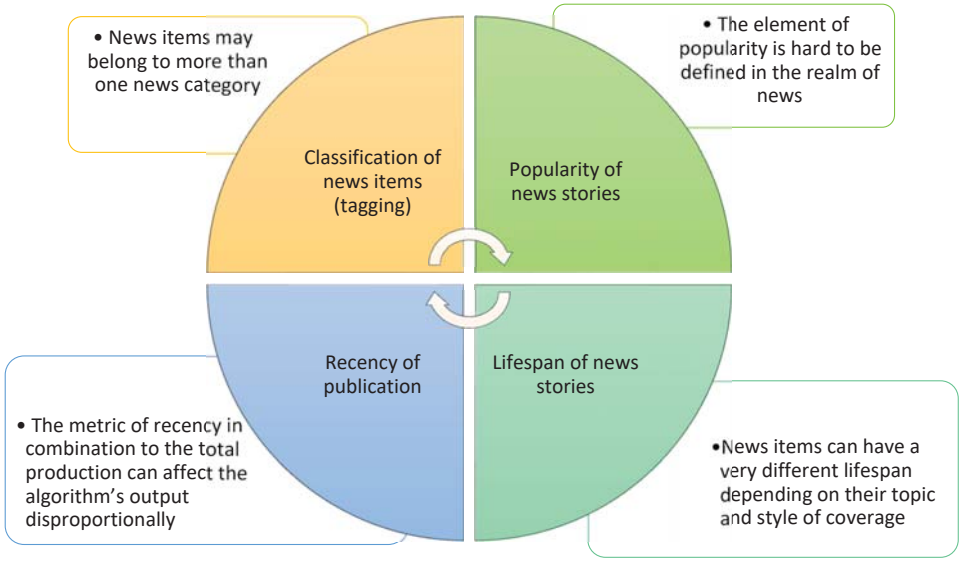


Figure 2. Design challenges for NRS related to the nature of news.

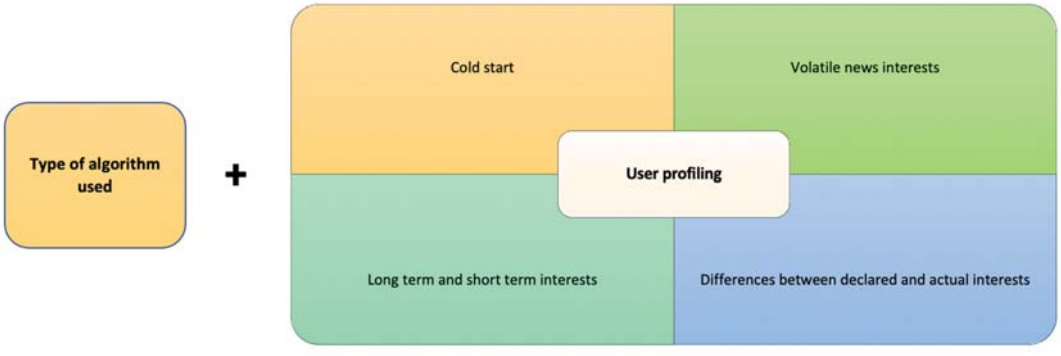


Figure 3. Challenges relating to user profiling in news recommender systems.

An important factor influencing how specific challenges will be treated pertains to the purpose of the recommender system as shaped by competing logics in the news organization. Smets et al. [58] argue that there is a crucial stage in the design of a recommender system in which the organization decides on the business and strategic goals they want to reach with the personalization service. A purpose-driven evaluation of recommender systems brings along the question of which stakeholder(s) are defining the recommendation purpose and how the conflicts and trade-offs between stakeholders are resolved and embedded in the system. More sophisticated algorithms not only combine hybrid logics in their filtering techniques but also include an element of surprise: serendipity [33]. The serendipity principle posits that the algorithm recommends items that are not only novel but also positively surprising for the user and propose a generic metric based on the concepts of unexpectedness and usefulness [32]. A key challenge of serendipitous recommendations is setting a balance between novelty and users' expectations. Serendipity is considered a quality factor for improving algorithmic output; it helps users keep an open window to new discoveries, it broadens the recommendation spectrum to avoid cases of users losing interest because the choice set is too uniform and narrow and helps integrate new items in order to acquire information on their appeal [49]. In addition to serendipity, the principles of novelty and diversity are deemed quality factors that can broaden the news menu and improve user perceptions and engagement [59]. Again, the 'right' level of novelty and diversity works as a trade-off for accuracy and can depend on the user's current situation and context [32]. Figure 4 depicts the main parameters for evaluating the quality of news recommendations. Recent initiatives aiming to run 'diversity-enhancing' algorithms focus on nudge-like personalization features, for instance, visuals to increase item salience, or item re-ranking in an attempt to curb rigid algorithmic recommendations [60]. Although nudging involves the steering of people in news paths that can enhance their news diet and knowledge [61], initiatives to nudge people toward diversity in their information exposure raise questions of autonomy and freedom of choice [62]. Non-transparent algorithmic nudging may undermine users' freedom of choice even if the principal objective (stimulating diversity) is a noble one [35,60].

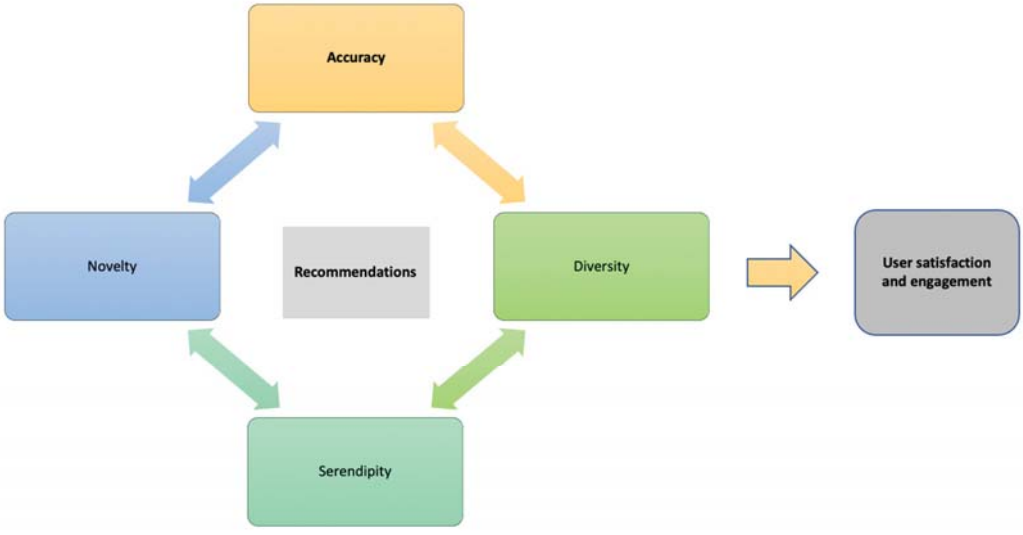


Figure 4. Parameters to evaluate the quality of news recommendations.

4. Experimental Setup for Validating an NRS Algorithm and Its Outcomes

The actual process of developing and rolling out news recommender systems within media companies remains largely under-researched [26]. In practice, explaining the workings of algorithmic systems is notoriously difficult [63] because of the opacity in their automated decision-making capabilities. Drawing on theoretical elements of algorithmic design [48,64], this study explores the implementation of a news recommender system within a leading news media organization in Cyprus, which started offering personalized news in January 2020.

The website is part of one of the largest media houses in Cyprus owning a television station, two radio stations, one newspaper and four established magazine titles. It is the leading online news player with 1.5 million users per month and 30 million page views monthly (source: Google Analytics). It covers political, economic and social affairs through a mainstream perspective. The website includes five sub-domains focusing on economy, sports, features, lifestyle and cooking, respectively. In addition to the website, the content is available on mobile apps. The website operates in a converged newsroom along with journalists from television, print and radio. Web metrics and analytics comprise an integral part of the outlet's content strategy. It manages the largest Facebook page for news in Cyprus with more than 140,000 followers and is also active on Twitter and Instagram.

The news recommender was developed as part of a research project. The outlet advertised its new service prompting users to register. Registration entailed basic user information, such as gender, age and interests. After visiting the website, registered users had the opportunity to log into MyNews, which was a webpage offering personalized stories. Each time, a user entered MyNews, a webpage containing 28–32 news items. These items appeared in a box-type layout, each one containing a photograph and the title of the news item. The boxes appeared in rows, each one displaying five articles.

The study draws on an experimental design employing the 'algorithmic audit' method [65] and more specifically the 'collaborative audit' which entails utilizing users as testers of algorithmic systems. The experiment was set up in a way to monitor the achieved accuracy of the algorithm and to probe the input parameters. To do so, users were divided into two groups: (a) *collaborating users* instructed to only view specific kinds of items (e.g., political ones) and (b) *ordinary users* who could elect freely what news item to engage with. The aim of dividing users into these two distinct groups was to monitor po-

tential accuracy differences stemming from users browsing behavior. Both groups involved registered users of the MyNews NRS service (see Figure 5).

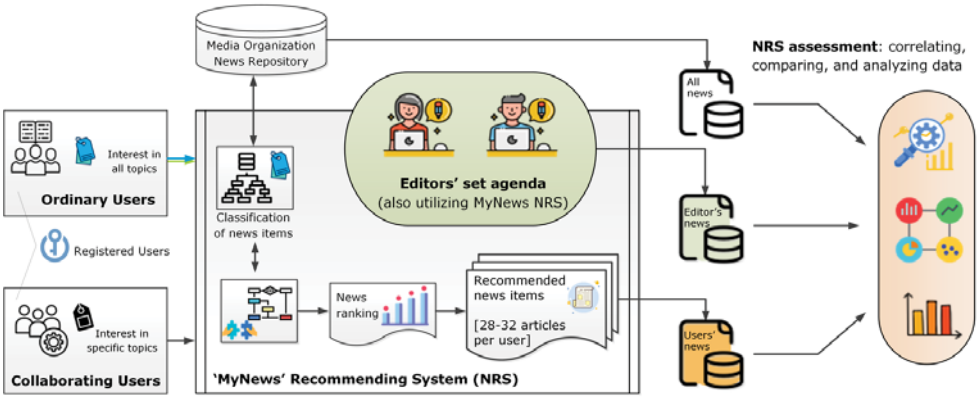


Figure 5. Visualization of the experimental approach: Design and datasets.

We posited the following research questions:

RQ1: When it comes to the news diet outputs produced by the editor-defined agenda and the algorithmic agenda, how different are they and in what ways?
RQ2: How effective is the recommender system in terms of accuracy? In other words, how likely is that the recommender system offers stories a user would be interested in, which is judged on the basis of the two distinct types of users?
RQ3: Are algorithmic recommendations more likely to lead to more clicks, i.e., does the recommender system increase users' engagement?

4.1. Participants

The study draws on the behavior of a total of 18 individuals who registered to the personalized news service of the outlet under study. These users were split into two groups. The first group comprised six individuals who were in essence collaborators of the researchers (*collaborating users*), each of whom was instructed to only click on news items that were of a specific type; e.g., one person was asked to only click on news relating to economy, while another was asked to consume only lifestyle-type news. More specifically, for these *collaborating users*, each was asked to read news pertaining to lifestyle, international news, local (Cyprus) news, politics, the economy and sports. The second group (*ordinary users*) included the remaining 12 individuals that were not given specific instructions and were asked to consume stories that were of interest to them. All users were additionally given the instruction to start perusing from the personalized MyNews area of the website.

4.2. Process and Data Collection

During a period of ten days (21 April 2020–1 May 2020), the following datapoints (see Table 1) were collected for each user and per session (each time they used the designated browser to visit the website under study).

Participants were fully informed that their activity was monitored in this manner; after providing their written consent to participate in the study, they were provided with an ad hoc created Chrome plugin pre-installed into a special browser to enable the collection of the aforementioned data. Users were asked to use this special browser any time they wanted to take part in this experiment (session). The plugin was transparently collecting and storing into a remote MongoDB database all the information necessary for analyzing the behavior of the news recommender algorithm implemented and applied by the medium. Data were only communicated and stored if users elected to press a specific button in their browser.

Table 1. Type of data collected.

	Type of Data	Abbreviation
1	the news items users clicked on	'clicked news items'
2	the news items contained in MyNews personalized area created for them by the recommender algorithm	'MyNews items'
3	the news items contained in the frontpage of the website at the time	'agenda items'
4	all news items published by the outlet per day	'pool items'

In addition to the aforementioned data, we collected information pertaining to the 'pool items' (the stories posted on the website) and also user information. Table 2 describes in detail the type of data collected. Having extracted the posts from the website, each entry was augmented with the user's information in order to associate users with content. User activity was captured by monitoring his/her click activity.

Table 2. Additional data collected pertaining to 'pool items' and user information.

	Pool Items	User Information
1	the section of the webpage where the post had appeared	the user's ID (usually an email address)
2	the news category assigned to the news item (as assigned by the medium)	the user's session ID (an auto-increment number)
3	the post's title	the session's timestamp
4	the post's URL	
5	the post's publication date	

Special mention needs to be made to the 'news category' variable mentioned in Table 2; this was an attribute assigned to each news item by the website, presumably the journalist responsible for writing the relevant article (i.e., these were not manually coded by the researchers). A total of 161 such categories were observed to be present on the news outlet under study; however, this included instances where mistakes had been made in tagging the article (e.g., the tags "Greece and "Greese") and duplications (e.g., using the tag "Greece" and "Ellada", a phonetic spelling out of "Greece" in the Greek language). When cleaned and grouped appropriately, a total of 33 different news categories emerged. Since however, using such a large number of categories would make the results difficult if not impossible to assess, it was decided to group these 33 categories into broader categories (e.g., "lifestyle news" were grouped together with "gossip"), ending with the nine categories reported in the results below (see Table 3). It should be mentioned, however, that this only affects the visual aspects of the results, as all calculations were made using the 33 original news categories assigned by the medium. Finally, it ought to be noted that the plugin maintained the order of posts as they appeared on the website, thereby ensuring that the order of the entries appearing into the database reflected the order the articles were read by the users. Every time a user accessed the website, the plugin collected all content and user-related information.

4.3. Datasets

In order to gauge the ability of the recommender algorithm to differentiate between users and respond to the research questions presented above, four types of datasets are necessary: a 'pool dataset', a 'MyNews dataset', an 'editor's agenda dataset' and a 'clicks dataset'.

Table 3. Relative frequencies of news items categories in the three datasets (percentage within dataset).

News Category	Pool Dataset	MyNews Dataset	Editor's Agenda
Politics	4.1%	5.9%	8.7%
Economy	15.3%	9.9%	6.2%
International	13.2%	18.4%	9.6%
Public Health	5.8%	2.6%	7.7%
Science and Technology	0.4%	0.1%	5.2%
Local (Cyprus)	12.5%	18.3%	10.6%
Greece	4.3%	6%	6.9%
Sports	32.3%	32.1%	12.2%
Lifestyle/Gossip	5.5%	3.8%	19.3%
Other Hard News (e.g., energy)	3.8%	2.6%	1.1%
Other Soft News (e.g., viral, culture)	2.6%	0.4%	7.5%
Reader Content	0.1%	0	5.1%

4.4. Pool Dataset

This dataset contained all news items published by the medium for each of the ten days of data collection, regardless of whether users interacted with them or not. The purpose of this dataset is to provide a baseline against which the output of the recommender system can be judged; if, e.g., a user was only interested in a category of news that appeared very seldomly on the website, it is natural for any algorithm to not be able to provide correct matches to this user's behavior.

The website published an average of 207.17 news items per day (median of 214). Table 3 presents the relative frequencies of the various news categories that the different news articles belonged to, from which it can be gauged that a plurality of the articles published belonged in the 'sports' category, which is followed closely by local (Cyprus-related) news, economy-related news and international news. All remaining news categories contained half or less of the aforementioned.

4.5. MyNews Dataset

This dataset contains all articles included in the personalized 'MyNews' area generated for each user and per session by the recommender algorithm. While merely by observing the relative frequencies of news categories in 'MyNews', one can reach early conclusions, e.g., that sports-related news is by far the most frequently observed category (32.3%) followed by some distance by local and international news, this might be deceptive; after all, such behavior could be indicative of the recommender algorithm successfully delivering sport-related content to users who are primarily interested in sports.

More importantly, a number of observations concerning the algorithm's behavior could be made on the basis of data other than the news category indicating designer choices. First, the algorithm consistently recommended between 28 and 32 articles to each user in each session. Secondly, there were no duplications of news items within any session, as could be expected. More interestingly, observing the times that the various items contained in 'MyNews' and the time a user session started, it became obvious that a number of rules have been set for the recommender system concerning recency: roughly 50% of the recommended items had been published within 3.5 h of the session's start, while 75% had been published within the last ten hours. The maximal time allowable lapsing between the start of a session and the publishing of a news item was 1439 min, which is one minute from 24 h. It then becomes obvious that the system had been designed to only deliver news appearing during the last day. However, this is the only hard rule that can be safely deduced; there appears to be an additional bias toward more recent items, but it is impossible to pinpoint how this operates exactly.

4.6. Editor's Agenda Dataset

This dataset refers to news items that were contained in the frontpage of the website, which was consistently structured in the same manner to contain a number of political,

economy-related, sports-related items, etc. that always appeared in the same space on the website's frontpage. This dataset indicates the agenda of the editors making these choices.

4.7. Clicks Dataset

This dataset contains the news items clicked on by the aforementioned users participating in this experiment (and the relevant attributes). Users differed in both the amount of clicks they performed in the duration of the experiment and in the amount of sessions they engaged in. On average, users clicked a total 252.1 news items, although this is significantly skewed due to the instruction to *collaborating users* to perform a minimum of 25 clicks per session, the difference between *collaborating* and *ordinary users* being statistically significant (Wlicoxon W = 73, $p = 0.044$).

5. Results

5.1. Comparing the Editor's and MyNews Agendas

The first research question concerned the structure of the news collections (in terms of news categories) offered by the different areas of the website: the Editor's Agenda (the frontpage) determined by the editorial team, the personalized MyNews Agenda produced through the algorithm, and the total pool of articles produced by the media organization, from which the aforementioned two draw their content.

Even a quick perusal of Figure 6 that compares the relative frequencies (in percentage) of the categories that the various news items are assigned to suggests substantial differences between the two (omitted from the figure for easier visualization are the near-empty categories: "reader content", "other hard" and "other soft news").

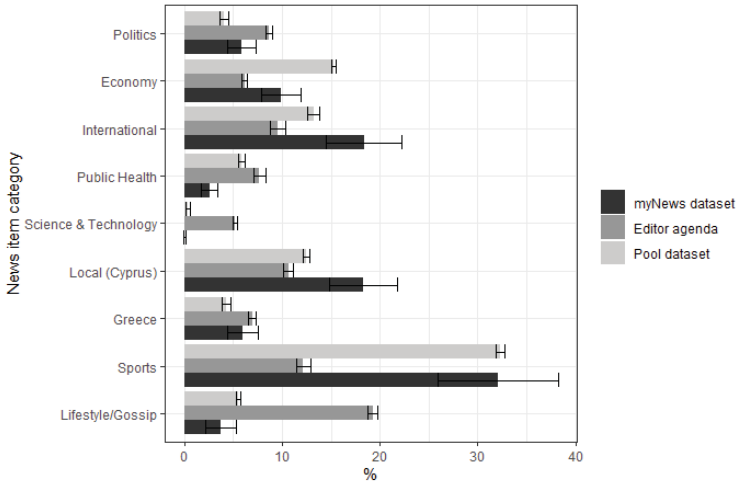


Figure 6. Relative frequencies of news items categories per dataset type.

The news environment produced by the two different processes, namely the editorial and algorithmic agendas, differ from each other. The editors have elected to follow a balanced approach with a mix of different news categories for their frontpage, with most categories being represented equally, covering between 5.2% and 12.2% of the available space. Interestingly, the exception to this rule concerns lifestyle and gossip-related items, which are overrepresented in the website's homepage (19.4%) particularly when contrasted with the relative amount produced in total (6.6% of all articles produced by the medium belong to this category). Other major categories of news (politics, international, local and sports) cover roughly the same large amount of space as would be expected from a mainstream news website aiming to cater for the diverse needs of a general audience.

To facilitate such comparisons, we take advantage of the fact that the overall pool of news items produced was collected; we divide the relative frequency (percentage) of each news category that appeared in the Editorial agenda (the frontpage) and in the MyNews area by the relative frequency of the categories of news items in the total pool of news items available for that session. If the result of this division (quotient) is 1, then the number of times a news item of the corresponding category in either the frontpage or the personalized MyNews section, is exactly as would be expected; were the two environments produced at random. Deviations from 1, on the other hand, suggest that either the algorithm or the editors who choose the frontpage items are favoring the specific news category (if the result of the quotient division is over 1) or that they are biased against the category (if the quotient is under 1).The boxplot (see Figure 7) examines exactly these ratios separately for the editorial and the MyNews agendas; the editorial agenda exhibits a soft news bias as shown from the overrepresentation of 'lifestyle and gossip' items. Similarly, news items tagged as relevant to 'Greece' and politics are over-represented, though much less so: about twice as many such items are contained on the website's frontpage (Editor agenda), as would have been expected by chance. Finally, concerning the editorial agenda, noteworthy is the larger than expected presence of 'public health'-related items, reflecting the ongoing COVID-19 pandemic. On the other hand, items dealing with the economy and international news are under-represented in the editorial agenda, as is the category of sports, although this is also the result of the large number of items in this category produced by the medium as a whole.

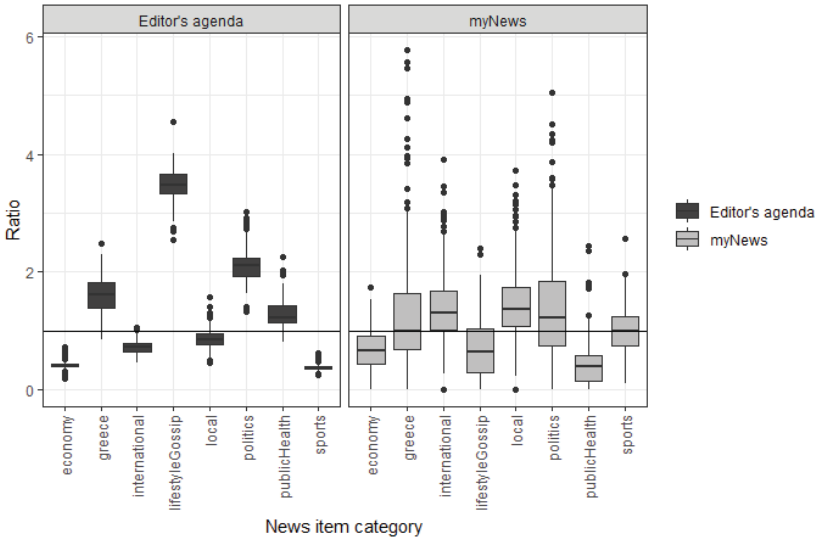


Figure 7. Boxplots of ratios of the Editor's Agenda and MyNews item categories to the available Pool items.

Concerning the MyNews area, we observe that the news stories offered are much closer to the distribution of categories in the total available pool of articles than the editorial agenda, with only economy- and public-health-related items being under-represented in the various MyNews agendas (personalized sessions) that different users in the sample viewed. However, it is not possible to reach any solid conclusions regarding an algorithmic bias from Figure 7, as it is produced on the basis of cumulative data from all users, who presumably have different interests—and the collaborating users among them, purposively so.

5.2. Evaluating Algorithmic Accuracy for Collaborating Users

While the aforementioned observations make it clear that there is a distinction between the news collection produced by the editors and the algorithm, they do not answer the second research question, i.e., whether the algorithm employed produces a distinct collection on the basis of deduced user interests. In order to examine this question, we need to take into account the type of user into our calculations.

We first focus only on data from *collaborating users*, who were instructed to only click on specific types of news (e.g., economy). While these users did not necessarily know the news category that the medium had assigned each item, it was hypothesized that they would be sufficiently accurate in clicking on only the 'right' (in accordance with the instruction they received) articles. An examination of the articles these users elected to click on suggests that this is indeed the case, since the relative frequency of clicking on the 'correct' type of article (i.e., where there was correspondence between instruction and the medium's assigned category) ranged between 78.8% and 92.7% of the articles viewed by these users (average 86%). These *collaborating users* then would be the easiest group for the algorithm to recommend, since they had highly discriminant behavior.

A dedicated examination of the MyNews/pool ratios (see Figure 7) for these users suggests that the algorithm indeed produced recommendations more likely to be clicked on by these users, with 'favored categories' being over-represented by a median of 2.16 times in the users' MyNews collections compared to the available pool, while 'non-favored categories' were under-represented by roughly 25% on average. However, this should not be taken to imply that the recommender system produces accurate recommendations; given these users' pre-designated behavior, their 'favored' categories should be over-represented to a much larger extent than observed, particularly during later sessions, when enough data on their behavior had been collected.

Indeed, on the basis of these data, we can construct a square matrix with rows representing the news categories collaborating users were instructed to click on and columns representing the categories of news items delivered to them by the algorithm averaged across sessions (Figure 8).

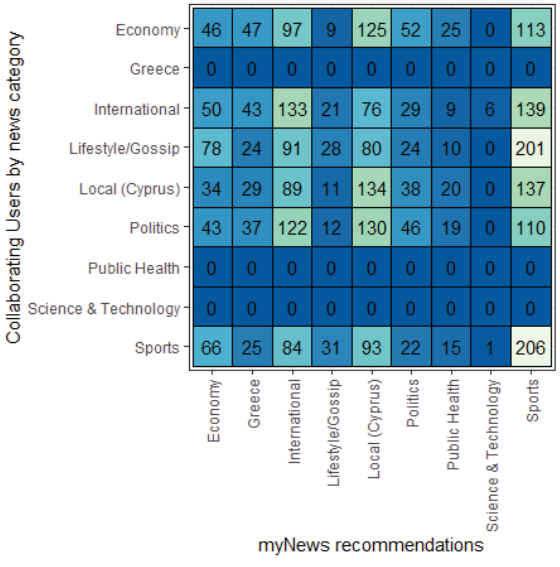


Figure 8. Matrix of frequencies of collaborating users.

This matrix can be re-composed to produce a per-news category confusion matrix for the dataset (Table 4), which can be used to produce various algorithm accuracy measures. The (unweighted) macro-averaged metrics for the algorithm was a precision of 0.211, recall of 0.19 and relevant F1-score of 0.2, confirming the aforementioned concerning the relative failure of the algorithm even for the collaborating users. The relevant overall accuracy metric was 0.747. In the calculation of these metrics, we omitted categories for which no collaborating user was designated ("Public Health", "Science and Technology" and "Greece"), since for these, no "True Positive" values could possibly exist.

Table 4. Re-composition of Figure 8 into a per-news category confusion matrix.

		True Positives	False Positives	False Negatives	True Negatives
News Category	Economy	46	271	468	2325
	Greece	0	205	0	2905
	International	133	483	373	2121
	Lifestyle	28	84	508	2490
	Local (Cyprus)	134	504	358	2114
	Politics	46	165	473	2426
	Public Health	0	98	0	3012
	Science and Technology	0	7	0	3103
	Sports	206	700	337	1867

The inability of the medium's algorithm to correctly include items that should have been recommended, at least for the relatively brief period of data collection, is also indicated by the lack of improvement in the algorithm model's accuracy score over time (Figure 9), as the flat linear trend indicates. So, while the algorithm did indeed differentiate between users, it did not create a news environment fully adapted to the assumed interests of these artificial, single-focused individuals. While this may indicate some faulty programming in the algorithm on the technical end, an alternative explanation can be offered. By design, the algorithm was required to select roughly 30 news items per session, which were additionally produced within the last 24 h (and more likely within 4 h of each session). However, the medium produced only, e.g., 11.3 and 16.4 politics- and lifestyle-related items as a whole per day. In other words, the algorithm failed to produce more politics-related items for the politics account, for the simple reason that it did not have enough such items available to pick from. The failure of the algorithm when it comes to the economy account is more perplexing, since roughly 41.8 such items are produced daily; the explanation for this might lie in the fact that these are mostly produced via the affiliated *EconomyToday* URL rather than the root website. It may be the case that the algorithm was designed to avoid cross-posting articles from such affiliated sites, although we have no means of ascertaining whether this is true on the basis of the data collected.

5.3. Evaluating Algorithmic Accuracy for Ordinary Users

Evaluating the algorithm's performance for the *ordinary users*, who received the instruction to choose whatever article they wished to view is less straightforward, since similar metrics cannot be calculated, as their interests are not known (as in the case of *collaborating users*). However, it is possible to construct a measure of distance between the items these users actually clicked on (indicating an overall 'profile') and the news collection presented to them by the algorithm in their MyNews area by calculating the distance between the relative frequency of articles in each category. Normalized, this index of distance takes values between 0 and 100, with larger numbers indicating greater distance between actual clicking behavior and the "MyNews" area. We can expect the algorithm to become better

at guessing users' behavior across time, since it has more data, and this should be reflected in smaller distances between the users' preferred environment (indicated by their clicking behavior in each session) and the algorithm's recommendations.

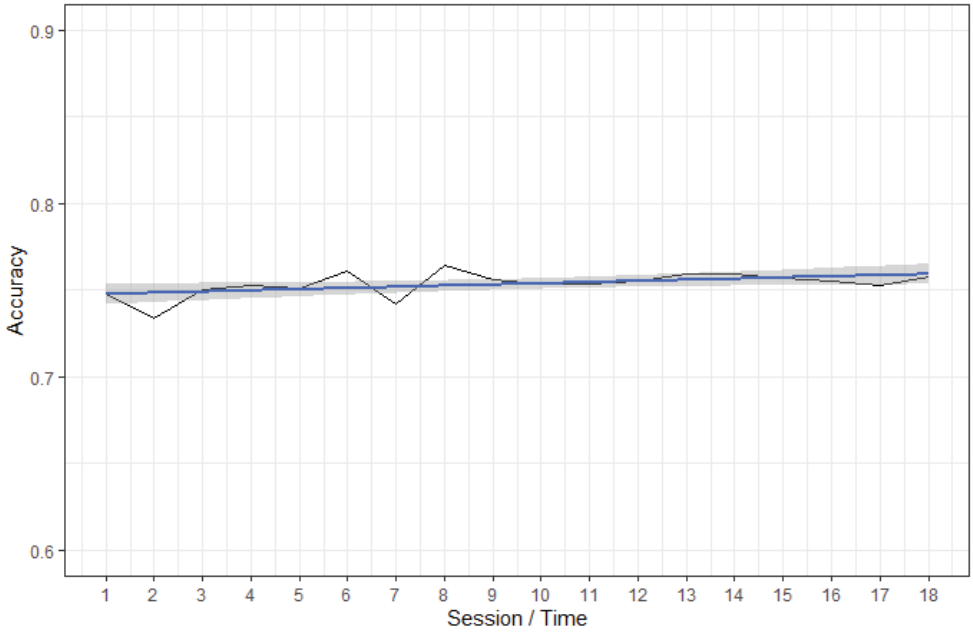


Figure 9. Accuracy score for *collaborating users* over time averaged across news type categories.

Figure 10 presents the linear trendline of exactly these distances across sessions for *ordinary users*, on average. While there is a decrease in the amount of distance between the pattern of clicking behavior and the news collection of the MyNews area, the decline is somewhat shallow with a decrease in distance of roughly 10% between the first and last sessions. However, the reader is reminded of the aforementioned design flaw in the algorithm (insufficient number of relevant articles in the pool the algorithm chooses from).

This relative absence of improvement in the algorithm's performance is also observable when considering the relative frequency with which *ordinary users* clicked on news items within their MyNews area. Results suggest that users did not choose news items from within the latter environment significantly more than they did at the beginning of the data collection period (Figure 10), suggesting a tentative negative answer to the third research question on whether the algorithm led to greater engagement.

5.4. Engagement over Time

After several sessions that would train the algorithm, one would expect that the algorithm would produce content that appealed to the users' interests and information needs, and therefore, the distance between media offerings in the MyNews area and user clicking behavior would diminish. Using the Euclidean distance metric, Figure 11 presents the polynomial trendlines of the distances across sessions for *collaborating users*, *ordinary users* and all users, on average starting from session 6 for each user. The failure of the algorithm to improve is apparent in the fact that these lines are relatively straight, whereas they would be expected to have a steep decline, although there is a small decline in distance following session 15, and the reader is reminded of the aforementioned design flaw in the algorithm (insufficient number of relevant articles in the pool the algorithm chooses from).

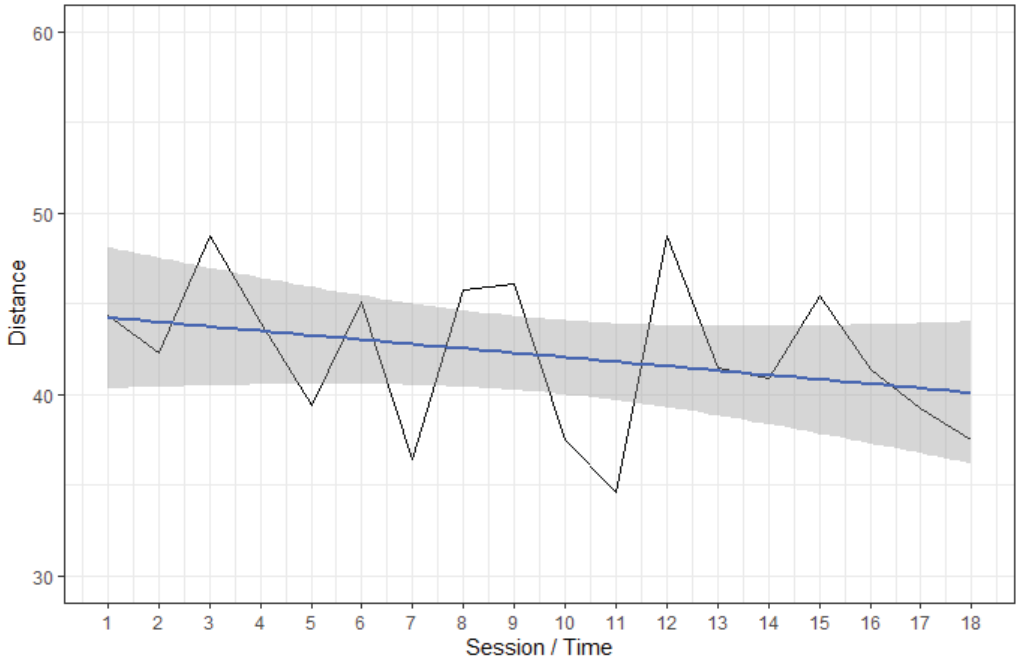


Figure 10. Distance between profile based ordinary user clicking behavior and MyNews outputs over time.

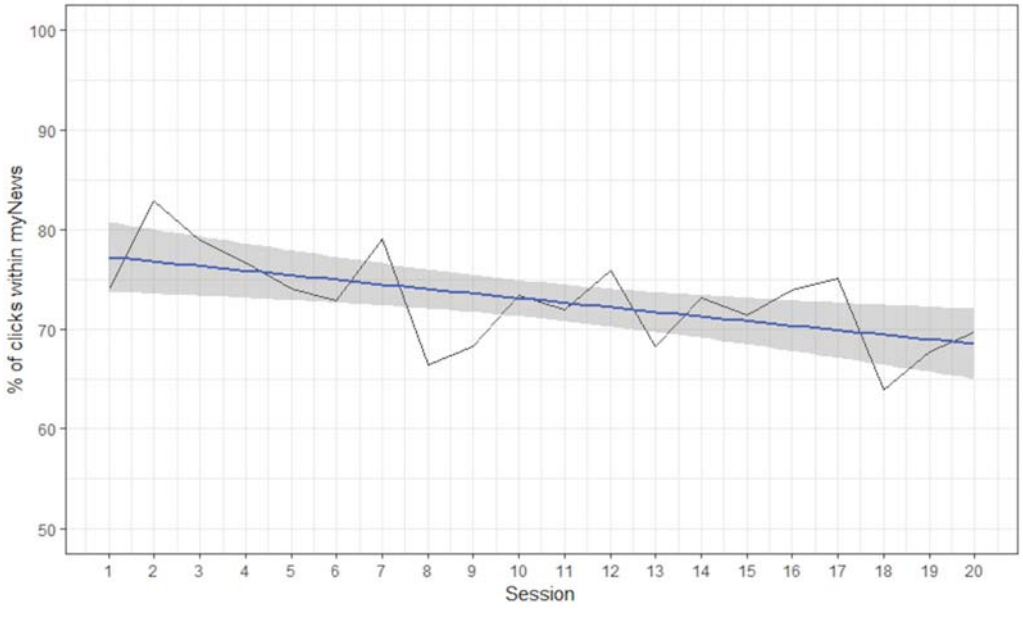


Figure 11. Percentage of clicks ordinary users performed within the MyNews area over time.

6. Conclusions

This work attempted first to synthesize relevant work and propose a generic news recommendation system which goes beyond accuracy and considers the civic role of journalism for an informed citizenry. Second, it attempted to examine the application of a recommender system to a national-level mainstream news organization in a small European country, which was a decision driven by the participation of the medium in a funded research project. Regarding the first research question exploring the main differences between the Editor's Agenda and the algorithmic agenda, the study shows that the Editor's Agenda was based on a balanced mix of hard and soft stories aiming to cater for the diverse needs of a general audience. On the other hand, a large number of recommendations offered in the MyNews sessions were significantly influenced by the available pool of stories produced on a daily basis. In other words, the news diet offered in the MyNews area is shaped by an algorithmic logic as opposed to an editorial logic identified in the Editor's Agenda [66]. Thorough examination of the data provided by users specifically instructed to only view certain types of articles (*collaborating users*) and *ordinary users* instructed to consume news stories on their own will and interests suggests that the application of the recommender system was problematic. While the algorithm does provide recommendations significantly different from what would be expected by chance, it ultimately fails to produce a personalized environment populated primarily by news items of the type that a user would be expected to view on the basis of their past behavior; this finding holds true for both *collaborating* and *ordinary users* examined here.

A careful examination of the instances in which the algorithm fails the most suggests that this is the result of design flaws rooted in problematic rules. Our findings provide empirical evidence showing [48] that unintended side effects of design decisions undermine the accuracy capacity of algorithms. These design flaws can be summarized as follows:

1. Voracity: We introduce the term voracity to refer to the large number of news stories expected to be offered in the MyNews area per user per session. Given the relatively small output of the news organization altogether, the identified under-performance of the algorithm was partly due to its being required to populate the personalized area with too many news items (28–32 stories).
2. Recency: The recency metric needs to be used with caution. In our case, it was mostly driven by the nature of the news content produced by the medium: timely, short-form stories aiming to build traffic. However, the requirement that all personalized recommendations must have been produced within a day of viewing—or preferably within the last four hours—significantly hampered the algorithm's capacity to provide accurate and relevant recommendations.
3. Unsystematic tagging: The classification and ranking of news items depends greatly on the tags assigned to news stories. Evidence of unsystematic tagging of articles had a negative impact on the algorithm's capacity to make accurate offerings.
4. Underuse of available content: Although the news portal under study is affiliated with four other websites, this content was excluded from the algorithm's repository. Considering the ambitious expectations of the MyNews area, designers and editors should provision a greater pool of content or limit the quantity of stories provided in the MyNews area.

Finally, the decreasing engagement of users with the MyNews area can be associated not only with the problematic levels of accuracy but also with the random filling of the 30-items sessions—shaped primarily by the availability of content—as opposed to a more sophisticated algorithmic provision of recommendations including the principles of diversity, serendipity and novelty [33].

Overall, the findings provide evidence that the effective design of news recommender systems depends not only on the particularities of the news domain as a media offering but also on the special traits and ideology of the news medium implementing the recommender system. By special traits and ideology, we refer to the characteristics of the news content produced, including (a) the quantity of news items produced, (b) the style of news reporting

(e.g., short stories satisfying the value of immediacy and clickability or explanatory stories having a larger life-span), and (c) the scope of the recommender system [58] (e.g., aiming to generate clicks and short-term engagement or provide a more balanced and diverse news diet [35,47]). To sum up, when implementing NRS, two major conclusions are drawn: First, design decisions need to be carefully associated with both the scope and the production capabilities of the news organization. Some metrics, for instance recency, are commonly used in NRS, but news organizations need to carefully define the metric according to their own needs and capacity. Second, all input metrics need to be validated as a whole and not separately.

This study does not come without limitations stemming primarily from the limited timespan of the experiment and the low number of participants. On the other hand, the heterogeneity of the media landscape [67] has produced a diverse set of online news media calling for the need to decode heterogeneous and emerging needs and thus types of NRS. Under this assumption, the utility of the present study lies in revealing significant insights about other like-minded models: namely, medium-sized mainstream online media focusing on short-form, current affairs-type of journalism.

7. Future Research

News recommender systems select, filter, and personalize news content, thereby shaping the news diet and knowledge level of citizens [68]; their algorithms make highly consequential decisions and thus exercise significant power [69]. At the same time, news media are tasked with relaying information to citizens, setting an agenda of common concern, acting as watchdogs to the powerful and providing an arena for public deliberation [70]. The overarching question therefore is how to ensure that news recommender systems enact the civic values of journalism in the direction of an informed citizenry while catering for the commercial needs of news organizations. Addressing this question in a principled fashion requires technical knowledge of recommender design and operation, and it also critically depends on insights from diverse fields [71], including journalism studies, psychology, policy and law. Following our discussion on diversity, serendipity and novelty, one critical question is how to operationalize these values and turn them into metrics specifically for the case for news offerings and also decide on the resulting trade-offs between competing values, stakeholders and overall scope. In this direction, further design approaches are needed to enable news recommender systems to conform to specific values while considering the capabilities and needs of different types of media.

Author Contributions: Conceptualization, P.S. and D.M.; Data curation, C.D.; Formal analysis, C.D.; Methodology, P.S. and D.M.; Writing–original draft, P.S.; Writing–review & editing, P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shearer, E. More than Eight-in-Ten Americans Get News from Digital Devices, Pew Research Center. Available online: https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/ (accessed on 21 January 2021).
2. Newman, N.; Fletcher, R.; Robertson, C.; Eddy, K.; Nielsen, R.K. Reuters Institute Digital News Report 2022. 2022. Available online: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf (accessed on 25 September 2022).
3. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2019. [CrossRef]
4. Myllylahti, M. Paying Attention to Attention: A Conceptual Framework for Studying News Reader Revenue Models Related to Platforms. *Digit. J.* **2020**, *8*, 567–575. [CrossRef]
5. Carlson, M. Confronting Measurable Journalism. *Digit. J.* **2018**, *6*, 406–417. [CrossRef]

6. Petre, C. *All the News That's Fit to Click: How Metrics Are Transforming the Work of Journalists*; Princeton University Press: Princeton, NJ, USA, 2021.
7. Nixon, B. The Business of News in the Attention Economy: Audience Labor and MediaNews Group's Efforts to Capitalize on News Consumption. *Journalism* **2020**, *21*, 73–94. [CrossRef]
8. Dimoulas, C. Machine Learning. In *The SAGE Encyclopedia of Surveillance, Security, and Privacy*; Arrigo, B., Ed.; Sage Publications Inc.: Thousand Oaks, CA, USA, 2018; pp. 591–592. [CrossRef]
9. Diakopoulos, N. *Automating the News: How Algorithms are Rewriting the Media*; Harvard University Press: Harvard, UK, 2019.
10. Negroponte, N. *Being Digital*; Alfred, A., Ed.; Knopf, Inc.: New York, NY, USA, 1995.
11. Thurman, N.; Lewis, S.; Kunert, J. Algorithms, Automation, and News. *Digit. J.* **2019**, *7*, 980–992. [CrossRef]
12. Thurman, N.; Schifferes, S. The Future of Personalization at News Websites. *J. Stud.* **2012**, *13*, 775–790. [CrossRef]
13. Van Drunen, M.-Z.; Helberger, N.; Bastian, M. Know your Algorithm: What Media Organizations Need to Explain to their Users about News Personalization. *Int. Data Priv. Law* **2019**, *9*, 220–235. [CrossRef]
14. Raza, S.; Ding, C. News Recommender System: A Review of Recent Progress, Challenges, and Opportunities. *Artif. Intell. Rev.* **2022**, *55*, 749–800. [CrossRef]
15. Bastian, M.; Makhortykh, M.; Harambam, J.; Van Drunen, M. Explanations of News Personalization across Countries and Media Types. *Internet Policy Rev.* **2020**, *9*, 220–235. [CrossRef]
16. Diakopoulos, N. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digit. J.* **2020**, *8*, 945–967. [CrossRef]
17. Picard, R. Twilight or New Dawn of Journalism? *J. Pract.* **2014**, *8*, 488–498. [CrossRef]
18. Bakker, P. Aggregation, Content Farms and Huffinization. *J. Pract.* **2012**, *6*, 627–637. [CrossRef]
19. Smyrnaios, N. *Internet Oligopoly: The Corporate Takeover of Our Digital World*; Emerald Publishing: Bingley, UK, 2018.
20. Rashidian, N.; Brown, P.; Hansen, E.; Bell, E.; Albright, J. Friend & Foe: The Platform Press at the Heart of Journalism, Tow Center for Digital Journalism, A Tow/Knight Report, 2018. Available online: https://www.cjr.org/tow_center_reports/the-platform-press-at-the-heart-of-journalism.php (accessed on 25 September 2022).
21. Spyridou, L.-P.; Veglis, A. Sustainable Online News Projects: Redefining Production Norms and Practices. In *Redefining Multiplatform Digital Scenario and Added Value Networks in Media Business and Policy*; Vukanovic, Z., Powers, A., Tsourvakas, G., Faustino, P., Eds.; Media XXI Publishing: Lisbon, Portugal, 2015; pp. 63–83.
22. Smyrnaios, N.; Rebillard, F. How infomediation platforms took over the news: A longitudinal perspective. *Political Econ. Commun.* **2019**, *7*, 30–50.
23. Rashidian, N.; Tsiveriotis, G.; Brown, P.; Bell, E.; Hartsone, A. Platforms and Publishers—The End of an Era, Tow Center for Digital Journalism, A Tow/Knight Report. 2020. Available online: https://academiccommons.columbia.edu/doi/10.7916/d8-sc1s-2j58 (accessed on 25 September 2022).
24. Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. [CrossRef]
25. Vryzas, N.; Vrysis, L.; Dimoulas, C. Audiovisual speaker indexing for Web-TV Automations. *Expert Syst. Appl.* **2021**, *186*, 115833. [CrossRef]
26. Bodó, B. Selling News to Audiences—A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media. *Digit. J.* **2019**, *7*, 1054–1075. [CrossRef]
27. Coenen, A. How The New York Times is Experimenting with Recommendation Algorithms, New York Times Open, 19 October 2019. Available online: https://open.nytimes.com/how-the-new-york-times-is-experimenting-with-recommendation-algorithms-562f78624d26 (accessed on 25 September 2022).
28. Jannach, D.; Zanker, M.; Felfernig, A.; Friedrich, G. *Recommender Systems: An Introduction*; Cambridge University Press: Cambridge, MA, USA, 2010.
29. Plattner, T. Why Personalization Will be the Next Revolution in the News Industry. Medium, December 15 2017. Available online: https://medium.com/jsk-class-of-2018/personalization-3a4cf928a875 (accessed on 25 September 2022).
30. Broussard, M. *Artificial Unintelligence: How Computers Misunderstand the World*; MIT Press: Cambridge, MA, USA, 2018.
31. Palomo, B.; Heravi, B.; Masip, P. Horizon 2030 in Journalism: APredictable Future Starring AI? In *Total Journalism*; Vázquez-Herrero, J., Silva-Rodríguez, A., Negreira-Rey, M.C., Toural-Bran, C., López-García, X., Eds.; Springer: Berlin/Heidelberg, Germany, 2022.
32. Karimi, M.; Jannach, D.; Jugovac, M. News recommender systems—Survey and roads ahead. *Inf. Process. Manag.* **2018**, *54*, 1023–1227. [CrossRef]
33. Møller, L.A. Recommended for You: How newspapers normalise algorithmic news recommendation to fit their gatekeeping role. *Journal. Stud.* **2022**, *23*, 800–817. [CrossRef]
34. Fenton, N. Left out? Digital Media, Radical Politics and Social Change. *Inf. Commun. Soc.* **2016**, *19*, 346–361. [CrossRef]
35. Helberger, N.; Karppinen, K.; D'Acunto, L. Exposure diversity as a design principle for recommender systems. *Information. Commun. Soc.* **2018**, *21*, 191–207. [CrossRef]
36. Bernstein, A.; De Vreese, C.; Helberger, N.; Schulz, W.; Zweig, K.; Helberger, N.; Zueger, T. Diversity in news recommendation—Manifesto from Dagstuhl Perspectives Workshop 19482. *DagMan* **2021**, *9*, 43–61. [CrossRef]

37. Hendrickx, J.; Smets, A.; Ballon, P. News Recommender Systems and News Diversity, Two of a Kind? A Case Study from a Small Media Market. *Journal. Media* **2021**, *2*, 515–528. [CrossRef]
38. Helberger, N.; Bernstein, A.; Schulz, W.; De Vreese, C. Challenging Rabbit Holes: Towards more Diversity in News Recommendation Systems. Media@LSE. Available online: https://blogs.lse.ac.uk/medialse/2020/07/02/challengingrabbit-holes-towards-more-diversity-in-news-recommendation-systems/ (accessed on 25 September 2022).
39. Pariser, E. *The Filter Bubble: What the Internet is Hiding from You*; Viking: New York, NY, USA, 2011.
40. Lafrance, A. The Power of Personalization, Nieman Reports, Fall 2017. Available online: https://niemanreports.org/articles/the-power-of-personalization/ (accessed on 25 September 2022).
41. Gillespie, T. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*; Gillespie, T.P.J., Boczkowski, K.A.F., Eds.; The MIT Press: Cambridge, MA, USA, 2014; pp. 167–193.
42. Just, N.; Latzer, M. Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet. *Media Cult. Soc.* **2017**, *39*, 238–258. [CrossRef]
43. Dylko, I.; Dolgov, I.; Hoffman, W.; Eckhart, N.; Molina, M.; Aaziz, O. The Dark Side of Technology: An Experimental Investigation of the Influence of Customizability Technology on Online Political Selective Exposure. *Comput. Hum. Behav.* **2017**, *73*, 181–190. [CrossRef]
44. Beam, M.A. Automating the News: How Personalized News Recommender System Design Choices Impact News Reception. *Commun. Res.* **2014**, *41*, 1019–1041. [CrossRef]
45. Dutton, W.H.; Reisdorf, B.C.; Dubois, E.; Blank, G. Search and Politics: The Uses and Impacts of Search in Britain, France, Germany, Italy, Poland, Spain, and the United States. Available online: https://ora.ox.ac.uk/objects/uuid:2cec8e9b-cce1-4339-9916-84715a62066c (accessed on 25 September 2022).
46. Bruns, A. *Are Filter Bubbles Real?* Polity Press: Cambridge, MA, USA, 2019.
47. Vrijenhoek, S.; Kaya, M.; Metoui, N.; Möller, J.; Odijk, D.; Helberger, N. Recommenders with a Mission: Assessing Diversity in News Recommendations. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, Canberra, Australia, 14–19 March 2021; The Association for Computing Machinery: New York, NY, USA; pp. 173–183. [CrossRef]
48. Diakopoulos, N. Algorithmic Accountability. *Digit. J.* **2015**, *3*, 398–415. [CrossRef]
49. Möller, J.; Trilling, D.; Helberger, N.; Van Es, B. Do not Blame it on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and their Impact on Content Diversity. *Inf. Commun. Soc.* **2018**, *21*, 959–977. [CrossRef]
50. Diakopoulos, N.; Koliska, M. Algorithmic Transparency in the News Media. *Digit. J.* **2016**, *5*, 809–828. [CrossRef]
51. Kompan, M.; Bieliková, M. Content-Based News Recommendation. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.299.4505&rep=rep1&type=pdf (accessed on 25 September 2022).
52. Bozdag, E. Bias in Algorithmic Filtering and Personalisation. *Ethics Inf. Technol.* **2013**, *15*, 209–227. [CrossRef]
53. Lu, Z.; Dou, Z.; Lianz, J.; Xiez, X.; Yang, O. Content-Based Collaborative Filtering for News Topic Recommendation. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Available online: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9611/9247 (accessed on 25 September 2022).
54. Darvishy, A.; Ibrahim, H.; Sidi, F.; Mustapha, A. *HYPNER: A Hybrid Approach for Personalized News Recommendation*; IEEE Access: Piscataway, NJ, USA, 2020.
55. Li, L.; Wang, D.D.; Zhu, S.; Li, T. Personalised news recommendation: A review and an experimental investigation. *J. Comput. Sci. Technol.* **2011**, *26*, 754–766. [CrossRef]
56. Feng, C.; Khan, M.; Rahman, A.U.; Ahmad, A. *News Recommendation Systems Accomplishments, Challenges & Future Directions*; IEEE Access: Piscataway, NJ, USA, 2020; Volume 8, pp. 16702–16725.
57. Lavie, T.; Sela, M.; Oppenheim, I.; Inbar, O.; Meyer, J. User Attitudes towards News Content Personalization. *Int. J. Hum.-Comput. Stud.* **2010**, *68*, 483–495. [CrossRef]
58. Smets, A.; Hendrickx, J.; Ballon, P. We're in This Together: A Multi-Stakeholder Approach for News Recommenders. *Digit. J.* **2022**, 1–19. [CrossRef]
59. Ekstrand, M.D.; Harper, F.M.; Willemsen, M.C.; Konstan, J.A. User perception of differences in recommender algorithms. In Proceedings of the 8th Conference on Recommender Systems, RecSys'14, Silicon Valley, CA, USA, 6–10 October 2014; pp. 161–168.
60. Reuver, R.; Mattis, N.; Sax, M.; Verberne, S.; Tintarev, N.; Helberger, N.; Atteveldt, W. Are We Human, or are We Users? The Role of Natural Language Processing in Human-centric News Recommenders that Nudge Users to Diverse Content. In Proceedings of the 1st Workshop on Nlp for Positive Impact, Online, 7 December 2022; pp. 47–59. [CrossRef]
61. Thaler, R.; Sunstein, C. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, 1st ed.; Yale University Press: New Haven, CT, USA, 2008.
62. Vermeulen, J. To Nudge or Not to Nudge: News Recommendation as a Tool to Achieve Online Media Pluralism. Available online: https://www.tandfonline.com/doi/abs/10.1080/21670811.2022.2026796 (accessed on 25 September 2022).
63. Pasquale, F. *The Black Box Society: The Secret Algorithms that Control Money and Information*; Harvard University Press: Cambridge, MA, USA, 2015.
64. Diakopoulos, N. Towards a Design Orientation on Algorithms and Automation in News Production. *Digit. Journal.* **2019**, *7*, 1180–1184. [CrossRef]

65. Sandvig, C.; Hamilton, K.; Karahalios, K.; Langbort, C. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, Paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a Preconference at the 64th Annual Meeting of the International Communication Association, Seattle, WA, USA, 22 May 2014. Available online: http://www-personal.umich.edu/~{}csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf (accessed on 25 September 2022).
66. Bucher, T. *If . . . then: Algorithmic Power and Politics*; Oxford University Press: Oxford, UK, 2018.
67. Splichal, S.; Dahlgren, P. Journalism between de-professionalisation and democratization. *Eur. J. Commun.* **2016**, *31*, 5–18. [CrossRef]
68. Moeller, J.; De Vreese, C. Spiral of Political Learning: The Reciprocal Relationship of News Media Use and Political Knowledge Among Adolescents. *Commun. Res.* **2019**, *46*, 1078–1094. [CrossRef]
69. Lundahl, O. Algorithmic meta-capital: Bourdieusian analysis of social power through algorithms in media consumption. *Inf. Commun. Soc.* **2022**, *25*, 1440–1455. [CrossRef]
70. Schudson, M. *Why Democrecies Need an Unlovable Press*; Polity: Cambridge, UK, 2008.
71. Stray, J.; Halevy, A.; Assar, P.; Hadfield-Menell, D.; Boutilier, C.; Ashar, A.; Beattie, L.; Ekstrand, M.; Leibowicz, C.; Moon Sehat, C.; et al. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *arXiv* **2022**, arXiv:2207.10192. [CrossRef]

future internet

MDPI

Article

Integrating Chatbot Media Automations in Professional Journalism: An Evaluation Framework

Efthimis Kotenidis [1,*], Nikolaos Vryzas [2], Andreas Veglis [1,*] and Charalampos Dimoulas [2]

[1] Media Informatics Lab, School of Journalism & Mass Communications, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece
[2] Laboratory of Electronic Media, School of Journalism & Mass Communications, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece
* Correspondence: efthimisko@gmail.com (E.K.); veglis@jour.auth.gr (A.V.)

Abstract: Interactivity has been a very sought-after feature in professional journalism ever since the media industry transitioned from print into the online space. Within this context, chatbots started to infiltrate the media sphere and provide news organizations with new and innovative ways to create and share their content, with an even larger emphasis on back-and-forth communication and news reporting personalization. The present research highlights two important factors that can determine the efficient integration of chatbots in professional journalism: the feasibility of chatbot programming by journalists without a background in computer science using coding-free platforms and the usability of the created chatbot agents for news reporting to the audience. This paper aims to review some of the most popular, coding-free chatbot creation platforms that are available to journalists today. To that end, a three-phase evaluation framework is introduced. First off, the interactivity features that they offer to media industry workers are evaluated using an appropriate metrics framework. Secondly, a two- part workshop is conducted where journalists use the aforementioned platforms to create their own chatbot news reporting agents with minimum training, and lastly, the created chatbots are evaluated by a larger audience concerning the usability and overall user experience.

Keywords: chatbot; media industry; news reporting; personalization; interactivity; journalistic practices; survey; workshop

Citation: Kotenidis, E.; Vryzas, N.; Veglis, A.; Dimoulas, C. Integrating Chatbot Media Automations in Professional Journalism: An Evaluation Framework. *Future Internet* **2022**, *14*, 343. https://doi.org/10.3390/fi14110343

Academic Editor: Wolf-Tilo Balke

Received: 29 September 2022
Accepted: 18 November 2022
Published: 21 November 2022

1. Introduction

The turn of the 20th century was characterized by many things, but perhaps one of the most important and influential changes was the introduction of information and communication technologies (ICT) that revolutionized many different facets of daily life [1]. A variety of professional fields got affected by the arrival of these technologies, and a prime example of that was the media and news industry, a sector always renowned for its unique entanglement with technological developments [2]. The way these new technologies were incorporated into the workflow of modern journalists were many and spanned across various fields of application, including the likes of content production and communication between the journalist and the audience [3]. Perhaps one of the most intriguing uses found for these new technologies was that of automated conversational agents. Commonly referred to as "chatbots", these programs were capable of communicating with users via the use of natural language [4], and it didn't take long for the media industry to realize the potentially massive benefits these digital tools could have in the process of creation, and particularly dissemination, of news content. Their rate of adoption varies based on a multitude of factors, but it is likely to increase with the rise of compatibility and the existence of more related expertise in the field [5]. In accordance with those parameters, chatbots in the field of journalism started as a relatively rare commodity, but as the technology matured, more and more media organizations started implementing them in their list of means to capture audience attention. As the development of conversational agents started becoming

more and more focused over the years, and chatbot technology reached its commercial stage, the wider accessibility of online creation platforms allowed the average journalists to involve themselves in this interactive procedure and create new and useful tools, even without the need for coding skills or specialized ICT knowledge.

Ever since the transition to WEB 2.0, there has been one attribute of this new and ever-evolving media landscape that has been considered particularly advantageous for its ability to better engage with audiences: interactivity [6]. Media organizations constantly seek more ways to captivate the attention of the public, and chatbots seem to fit that bill perfectly, seeing as they are a type of tool designed based on principles that center on communication and conversational intelligence [7]. To this end, this manuscript attempts to highlight and overview some of the most well-known and readily available contemporary digital tools that journalists have at their disposal for chatbot creation. At the same time, it aims to present an evaluation framework for said tools, which includes allocating them to certain categories based on the amount and type of interactivity that they offer, as well as judging their performance and accessibility in general. The selected programs are specifically targeted to not require any amount of prior coding experience on behalf of the journalist for them to fully function and create a finished product capable of interacting with audiences and establishing new communication channels.

The motivation behind this work stems from the general question of whether it is feasible to integrate chatbots in professional journalism and, if so, under what circumstances. This question can be broken down into whether the existing platforms are intuitive and ready to be used by professional journalists to create chatbots and whether those bots are helpful, useful, and engaging for a broader audience. To that extent, it is important to design a framework for their evaluation that can set the directives of how they can be improved upon. This paper considers some of the most important platforms that are being used right now and presents a three-fold evaluation of them, addressing the scientific questions that have been set.

Other than this introductory segment, the remainder of the paper is organized as follows: In Section 2, the related work on chatbots is presented, including a brief overview of the term, research on journalistic chatbots, and the role of interactivity. Section 3 introduces the proposed framework for integrating chatbots in professional journalism alongside the evaluation methodology that has been adopted. The selected creation tools are also presented there. In Section 4, the evaluation results are discussed, and finally, Section 5 summarizes the conclusions and the future research plans for this project.

2. Related Work

2.1. Chatbot Categories and Architecture

The term "chatbot" can be defined in several ways, and many of them include a wide variety of programs. For this study, we are going to define a chatbot as a software application that utilizes natural language to communicate with humans [8]. Such programs have a very wide variety of applications thanks to their flexibility, and they are systematically utilized by sectors like customer service for their ability to converse with humans in a relatively natural and meaningful manner [9].

When it comes to a taxonomy of the available chatbot models, one of the major ways in which conversational agents can be distinguished from one another is their architectural design. Specifically, there are two prevailing categories that all chatbots fall under, depending on what procedure they follow to respond to users: "retrieval-based" chatbots and "generative" chatbots [10]. The first category–that of retrieval-based chatbots–comprises programs designed to communicate with the user via predetermined responses [11]. Conversational agents that fall under this model operate by searching for a reply in a pre-established repository and serving it to the user according to their input [12]. The large majority of chatbots that can be found on the web today follow this model of operation due to its simpler structure compared to the alternatives [11]. However, even though the architecture of retrieval-based chatbots presents many advantages–especially for users that don't have

any coding experience or specific ICT knowledge–this method is subject to some limitations that can potentially restrict the scope of the final product. Those limitations mainly relate to the fact that the procedure of choosing between a set of already existing responses makes it very challenging to customize the chatbot for particular situations and relegates it to having a more "passive" role in conversations [12,13].

On the other end of the spectrum, the category of generative chatbots consists of software capable of creating new responses from scratch to better match the queries of the user with the help of techniques such as Machine Translation [14]. The result of this is an often fairly convincing effect, where the chatbot can uphold a conversation with a user in a very natural manner. This procedure is a lot more demanding than its retrieval-based counterpart since it requires a substantial training dataset to function properly, but it provides the significant benefit of being able to respond to user inputs for which predetermined responses don't exist, something that the retrieval-based model falls short of. At the same time, however, it is far more likely for it to exhibit major grammatical errors, or make other similar mistakes, compared to a solution that relies on retrieving responses from a designated archive [15].

Natural Language Processing has been advancing rapidly over the past years, and the utilization of new and innovative techniques, such as Deep Learning, suggests that generative models will be the future of chatbots moving forward [16]. However, the typically enormous datasets required for properly training such systems, as well as the complexity of text generation, which constitute two of their primary characteristics [15], dictate that retrieval-based solutions will remain the far more common alternative for the time being, especially when it comes to chatbot creators that are less knowledgeable in the programming department.

2.2. Chatbots and Media Automations in Journalism

Many people consider chatbots to be a fairly recent development, given the fact that commercial use of this technology has only started to become prominent in the past decade or so. In reality, however, efforts by researchers to create a program capable of conversing with humans naturally have been going on for the better part of a century. Often regarded as the predecessor of all chatbots, ELIZA was created by the German scientist Joseph Weizenbaum [17] and managed to engage its conversational partners to such a degree that many of them reported that they believed they were addressing a real person [18]. From that point onwards, many researchers attempted to simulate human communication with computer programs, which led to the creation of many prominent examples of capable conversational agents over the years.

Despite the rich historical development of chatbot technology, however, what indeed constitutes a relatively new phenomenon is the inclusion of these programs in the process of creating and disseminating news. This was a by-product of the general tendency for computational procedures to graduallywork themselves into the practices of the journalistic profession, subsequently creating what was later called "automated journalism", frequently also referred to as "robotic", "algorithmic", or "computational" journalism [19]. Chatbots in journalism can be used for efficient user interaction in the sense that they provide a more human-like way of navigating and accessing news-related content. Generally speaking, chatbots are a sub-section of the more general term AI agent, which is used to describe artificially intelligent software capable of performing a variety of tasks. In the case of rule-based chatbots, these tasks center around providing the user with content based on their answers to the predefined questions of a conversation. In a more sophisticated approach, chatbots can also generate original content, through algorithmic techniques like text generation, document summarization etc., by applying basic concepts of algorithmic journalism.

As far as chatbots are concerned, many decades had to come by until the first proper use of one for purely journalistic purposes could be identified. Specifically, in 2014, the Los Angeles Times employed the services of an automated AI agent that was capable of data extraction and simplified content creation. The program, aptly named "Quakebot", was

tasked with monitoring data from the U.S. Geological Survey and utilizing any information it could find regarding seismic activity to write and publish simple reports in realtime. Even though Quakebot only amounted to a very simple use of this type of technology–and even at the time, it was being compared to nothing more than an intern with a lot of time in their hands [20]—it still embodied the essence of what the introduction of such programs meant for the journalistic profession: cheap and easy to maintain labor, that could substitute, or even surpass, human journalists in some tedious and time-consuming tasks.

On that basis, what followed during the next half of the decade was a massive increase in the introduction of automated elements into journalism, with chatbots being one of the technologies at the forefront of that wave of change, as, over time, many news organizations started realizing the potential benefits of automatic content dissemination [21]. Companies gradually started utilizing the unique qualities of conversational agents to spread their content more efficiently-particularly through websites like social media–which led to the creation of the term "News Bots", a sub-category of chatbots that specialize in interacting with audiences and spreading news information [22]. The inclusion of these automated programs in the distribution of news content made the whole process much more efficient and allowed audiences to interact with news organizations in new and more engaging ways that proved to be very effective, thanks in no small part to the interactivity and personalization they offered [23].

At the same time, following the example of Quakebot, chatbots capable of data extraction and content creation started becoming more prevalent as the news industry attempted to adjust to the rapidly evolving media landscape. Faced with the ever-increasing load of information available on the internet, journalists started using AI agents capable of sifting through large amounts of data and collecting relevant results for their work in a process called "Data Mining". This procedure was able to help media staff identify stories that have editorial value, according to the parameters set by the journalist, as well as provide aid in more specific tasks such as information verification and event monitoring [24]. Similarly, automated content production proved to be yet another way to take advantage of these versatile programs in the realm of AI-assisted news. Chatbots and related algorithms were developed to produce content on their own, with little to no human intervention. Machine-generated news content is often powered by artificial intelligence or similar machine learning algorithms [25], and it has been one of the more defining factors of the journalistic profession in the past few years [26], as it has led to many upsets in the industry, provoking many researchers, as well as practitioners into questioning whether or not these programs could potentially prove to be a threat for industry workers [27,28]. The prevalence of these new digital tools is in part responsible for the cultivation of a work environment in which digital literacy is one of the most important aspects, with the imminent re-defining of journalistic skill sets coming to the forefront [28,29]. Despite that controversy, however, what remains an undeniable fact is that chatbots and other AI agents like news writing algorithms are seeing extensive use in news media production today, with many industry-leading organizations like Forbes and the New York Times utilizing them as content creators, with the final product being almost impossible to distinguish from human writing [30]. The combination of these abilities exhibited by automated software–to comb vast amounts of data and then morph the relevant information they find into a compelling narrative–has already expanded the news writing universe in a major way, and they will, in all likelihood, continue to do so as the technology improves [31], further fueling the domain of media automation in professional journalism, which can also augment the interaction through the incorporation of media agents [32].

All of the chatbot categories mentioned above have seen extensive use in the field of journalism over the past decade. Among them, however, the category of news dissemination stands out as the field of application that incorporates chatbots the most [23]. These types of chatbots are also the most relevant ones to be examined for this study, as they are designed with the explicit goal of improving audience engagement and introducing more interactivity into news distribution [33]. This practice has transformed how media is shared

and consumed since it was adopted by many prominent news organizations such as CNN, The Guardian, and The Washington Post, all of which created their own versions of news-sharing chatbots by the end of 2016, which was a year that saw a particularly large surge in chatbot creation in general [34]. Figure 1 reflects the current perspective of chatbots in the context of media automation in journalism in relation to the evaluation framework, which will be proposed later in the study. More specifically, this conceptual diagram highlights the two-way relationship that is present between the individual processes in journalism and how this feedback loop is utilized in order to reach the concessions needed for the proposed evaluation framework. In turn, the evaluation framework also feeds into this information cycle in its own way by providing a way to accurately assess the usability and usefulness of relevant tools.

Figure 1. A conceptual diagram regarding chatbot media automation in professional journalism and the proposed evaluation framework [28].

Overall, a variety of literature exists in the realm of chatbot usage, specifically for journalistic purposes. Researchers over the years have covered a wide spectrum of chatbot applications in the field and have tackled issues like their usage in automated news dissemination [11,22,35], information gathering [34], user newsfeed personalization [23], as well as the use of conversational agents to establish a multi-media approach in the news [33,36], or as a means of studying the relationship between the journalist and the audience [37] as well as the cross-cultural social context those bots are being employed in [38]. Despite that, however, the majority of this research–with only a few notable exceptions–seems to focus more on the potential practical uses of these programs and less on a valid framework with which chatbot creation tools can be evaluated. Even though conversational agents are undoubtedly very useful in the media sphere, it is important to realize that the average journalist is often unfamiliar with the more technical aspects of chatbot creation. To that end, there exists a distinct lack of research focusing on evaluating easy-to-access tools for the average worker in the media industry, a flaw that this paper aims to remedy by prioritizing the opinions of individuals who are familiar with the field of journalism and integrating them via a practical use-case scenario.

2.3. Interactivity in Chatbot Design

Interactivity in media is a concept that can be defined in many different ways, depending on the angle through which one approaches it. Different parties have attempted to provide different definitions for it, depending on whether the interaction is an observer between people, between a user and a machine, or even based on the ability of a user to control and alter a given message [39]. Perhaps the most all-encompassing definition for the term, however, was given by Liu and Shrum [40], who describe it as "The degree to which two or more communication parties can act on each other, on the communication medium, and on the messages and the degree to which such influences are synchronized". This admittedly broad definition manages to encapsulate every relevant aspect of interactivity when it comes to its applications in online media and lends itself well to the examination of programs like chatbots within the medium. After all, a prerequisite for interactivity is the multi-directional flow of information between the user and the producer of the content,

where the audience is the recipient of information but also provides some sort of feedback on it [41].

Interactivity is generally regarded as the most defining factor of new media, and oftentimes it can even be considered a necessity for them [42]. Based on that, media companies started to focus more and more on interactive features for their content ever since the internet became paramount for the news industry [43] by introducing several different features that were impossible to include in traditional media. Those features are implemented because they provide numerous benefits, like human-to-human interaction (or, in some cases–like the ones we will examine later–human-like interaction), emphasize socialization, and–other than audience engagement–they can also help in the creation of online communities of readers [44] which is an added benefit, associated with even more positive outcomes for media organizations. Generally, interactivity seems to be a driving force behind longer audience engagement in media, which is not at all surprising, given the fact that it has also been proven to significantly boost user engagement, even throughout vastly different mediums like public displays [45] and digital games [46]. There is also a strong association between the need of the audience for entertainment and the use of medium interactive features to fulfill that need [44].

Many researchers have suggested models with which interactivity can be measured or segmented into different types depending on a variety of factors. One such model for categorizing interactivity that is particularly relevant to this study was proposed by Jensen [47], whose definition of interactivity is similar to the one given in this chapter, "the measure of a media's potential ability to let the user exert an influence on the content and/or form of the mediated communication". Based on that definition, four sub-types of interactivity are proposed: Transmissional, Consultational, Conversational, and Registrational interactivity. The first one, Transmisional interactivity, refers to a medium's ability to allow the user to choose something out of a constant stream of content in a system where there is no two-way communication, like a Teletext service, for example. Consultational interactivity is similar but requires the user to be able to choose something on demand, out of an already pre-produced library of information, like an online encyclopedia, or a streaming service, which suggests the existence of a return channel. The third one, Conversational interactivity, applies to systems that allow the user to create and input their own information in a two-way media channel like an e-mail client. Finally, Registrational interactivity refers to a medium's ability to register information from a user and adapt or respond to it. This applies to more "intelligent" systems that can answer specific requests and cater to a user's needs. It is worth noting that–even though this isn't explicitly mentioned by Jensen in his original paper [47]–this segregation seems to be presented in ascending order of sorts, with each subsequent category allowing more and more liberties to the user, and providing them with more freedom to influence the outcome of the interaction. As part of the proposed evaluation framework, this paper is going to attempt to categorize some of the most notable examples of chatbot creation platforms into the above categories, as well as take a more in-depth look into their specific characteristics to distinguish between them in a manner that will be further elaborated below, using a three-way evaluation approach.

3. Integrating Chatbots in Professional Journalism

3.1. A Framework for Journalism-Oriented Chatbots

As stated in previous sections, journalism is already intertwined with chatbot usage to a significant degree. Media organizations, as well as journalists as individuals, routinely resort to chatbots as a tool for their professional needs. There are still, however, some open questions concerning the integration of chatbots in journalism. These refer to the use cases where chatbots can improve news reporting, the relationship between an organization and the audience, and also how capable, confident and engaged journalists and the audience feel when it comes to using them. For professional journalists, this means programming chatbots to bring their work forward, and for the audience, it means using a chatbot to navigate news content. Even though there is no question that news dissemination is by far

the most widely used application for chatbots in the field of media, one would be correct to assume that these versatile programs can also be utilized for other, more unique scenarios. For example, the interactive platform provided by conversational agents can be used as a "news assistant" of sorts by transferring additional information to the reader regarding the topic of interest when they request it. This approach has been examined by embedding chatbots within news articles [33], and similarly, these applications can also be utilized in other creative ways, such as gathering information from the public [34]. It is possible to envision the usage of chatbots for other creative work as well, although in most cases, more advanced conversational intelligence would be required in order for these programs to present a compelling case for their usage in these scenarios.

In this manuscript, we focus on how this symbiotic relationship between chatbot and journalist can be improved upon by examining some of the most accessible alternatives for chatbot creation aimed at workers in the media sphere that don't possess any programming knowledge or any other similar expertise in the field of ICT. Given the above, to design a use case scenario for integrating approachable chatbot building in professional journalism, we need to account for two discrete aspects of chatbot usage in journalism: Front-end usage, also known as the part of a program the user interacts with, as well as Back-end usage, which refers to the part that the designer of a program interacts with.

The frontend refers to the interface of the end product–in this case, the finalized News Bot- and it is targeted at news consumers. This is the part of the program that the average user will be interacting with to receive news updates, as it will be explained in later sections. For the purpose of the front-end evaluation, we will be examining how intuitive the end product is, how much it helps the average news consumer, and how interactive and helpful the chatbot is perceived overall.

When it comes to the backend refers to the chatbot creation platforms themselves. These are going to play the role of a mediator in the chatbot creation process, as the journalist will be interacting with each of them to build a chatbot from scratch. For this part of the evaluation, we focus on the ease of use when it comes to creating a new chatbot, the interactivity features that are included in each platform, as well as all the monitoring and quality of life features that are presented in the journalist upon creation of the program. The usefulness of each platform to the journalist is also taken into account. With all of that in mind, the following use case scenario was developed by the researchers for the subsequent evaluation of the examined tools and platforms.

Use Case Scenario

This use case scenario aims to provide context as to how the chatbot creation platforms included in this manuscript will be evaluated. It does so by highlighting a practical, real-world example from the perspectives of the main stakeholders in the procedure while aiming to be both realistic and clearly understood.

The primary actor in this scenario is a professional journalist that works in the media industry on behalf of a news organization. This person will be the one interfacing with the back end of the chatbot construction platforms to create the desired product. The specific type of journalist represented in this scenario is well accustomed to digital technologies and has at least a basic understanding of current information and communication technologies. Having said that, however, no programming skills or any other specialized ICT knowledge is assumed, as the chatbot creation platforms have been specifically chosen to provide a coding-free experience and can be used by any professional willing to invest time in them. The primary motivation of this actor is to create a product that will adequately serve the audience of their news organization and increase user engagement.

On the other end of the spectrum, the secondary actor is the end-user of the product, which in this case is the news consumer. This person will be the one receiving news content by interacting with the product created by the journalist. Similar to the primary actor in this scenario, this type of user has at least a basic understanding of current technologies

and chooses to primarily receive their news online. The basic motivation of this actor is to consume news articles in an easy and digestible way.

The systems used in this particular use case scenario are the three chosen chatbot creation platforms that are being evaluated, which will be presented in the following section. The journalist, as the primary actor, is tasked with creating a chatbot on behalf of their organization to better disseminate news content. To create a successful product for journalistic use that accommodates the needs of the presented research, this chatbot needs to align with certain parameters. For this study, it was determined that the final product needs to be able to disseminate news content:

- Automatically;
- Systematically;
- Interactively.

The automation of the news dissemination procedure is the reason why chatbots are being used in the first place. In addition to that, however, the requested application needs to be able to prescribe certain characteristics to the news consumption process by actively making it more engaging. For this reason, the journalist needs to be able to utilize the environment of the chatbot creation platforms to accomplish these goals in the most accessible way possible. To that end, parameters like the complexity of available features, the overall responsiveness as well as the ease of use were chosen as some of the most important variables to be taken into account for the final evaluation.

For this work, we assume that the end-user will be contacted by the chatbot in intervals that range from a single day to an entire week, as per their subscription preferences. In that scenario, the chatbot creation platform needs to be able to provide an end product capable of engaging with the user systematically and interactively without overwhelming them while keeping the entire procedure as simple and fluent as possible. Additional options, like features that accommodate back-and-forth communication between the primary and secondary actors (the journalist and the news consumer), are also taken into account.

3.2. Methodology for the Evaluation of Chatbot Platforms for Journalism

As already explained, a big aspect of the motivation of this work is to analyze the feasibility of integrating chatbots into professional journalism. To answer this, it is important to analyze whether journalists can design chatbots without prior training and whether the audience feels confident and engaged in using the frontend. For this reason, we have taken into consideration three important platforms that were evaluated in the aforementioned directions. In the next sections, the main characteristics of the platforms, as well as the evaluation experiments, are presented and explained.

3.2.1. Chatbot Creation Platforms

Chatbots, like most technologies in their infancy, started entering the commercial space slowly and experimentally. Over the years, however, the market started becoming more and more saturated with a variety of platforms capable of chatbot creation. The ability to utilize those programs for easier news dissemination, among other things, presented many benefits for media organizations, as discussed in earlier chapters, and thus it was adopted fairly quickly, to the point of becoming an industry standard within a few years. Specifically, a large surge in chatbot integration was observed during 2016, with numerous media companies announcing their implementations within the span of a few months [34]. This wave of chatbot innovation can be partially attributed to social media, specifically Facebook's decision to open up its ecosystem to developers by natively supporting chatbots through its messaging service. This marks a turning point for the media landscape, not only for the news organizations themselves but also for the consumers, as the wider availability of chatbots also plays a major role in the democratization of certain services since these programs can be made available to a very large number of users via platforms like social media and messaging applications [48]. Nowadays, online tools exist that allow the average worker in the media industry to take advantage of their coding-free environment and create

a fully functional conversational bot for their own journalistic purposes. This approach has even been adopted by many prominent media organizations that decide to utilize these platforms for their chatbot needs. Some of these companies, like CNN, preferred to outsource the creation of their bot to a third party, in this instance, a chatbot-building company named SPECTRM. Many other industry giants, however, decided to try their hands on these platforms and experiment with what those online tools could offer them by allowing their journalists to create their own version of a chatbot for use in the newsroom.

It is important to note that none of the currently existing online creation platforms were explicitly developed to build chatbots suitable for journalistic work. Most of the software platforms available online cater to a generalist audience and, in many cases, specialize more by offering extra features for some sectors that use chatbots very extensively, like customer service and marketing. In that context, the media industry has been trying to take advantage of the already existing tools to fulfill the needs of the audience by adapting to their limitations and nuances. Since this paper aims to evaluate the most readily available online tools that can be used by the average worker in the media industry, the criteria with which the choice of platforms was made are as follows.

First off, the creation platform should include a user-friendly environment that doesn't require any coding expertise or comprehensive ICT knowledge on behalf of the journalist. The reason behind this, when it comes to the scope of the paper, is the need to identify the tools that can appeal to the widest possible audience while taking into account that many media industry workers are not entirely familiar with concepts like programming and creating an application for the general public. In addition to that, however, the accessibility of the user interface plays a particularly important role, as the chosen platforms will be evaluated–in part–via the use of a workshop, as will be explained in detail in the upcoming sections. This narrows the list of selected platforms down to options that lend themselves to be easily presented and taught in a workshop setting, where the participants can follow through with creating their chatbot before being asked to evaluate the user experience.

Additionally, the selected platforms must offer a usable free plan that journalists can utilize to create a fully functional bot. Most companies in the field of chatbot creation operate under a hybrid business model, where they offer a free plan with limited functionality and a premium one with more features (Business models for the platforms used in the study can be found here: (1) https://quriobot.com/pricing (accessed on 28 September 2022) (2) https://snatchbot.me/pricing (accessed on 28 September 2022) (3) https://chatfuel.com/pricing (accessed on 28 September 2022)). We are mostly interested in tools that allow users the freedom to begin experimenting with chatbot creation without any commitments, and thus only platforms with a usable free plan were taken into account. Finally, the chosen tools must exhibit characteristics that align with usage in a journalistic environment. This includes a list of criteria that will be further elaborated on in the evaluation section. Based on the above, the three following chatbot-building tools were chosen and will be examined below: Quriobot, SnatchBot, and Chatfuel. What follows is a short description of the way each one of the selected platforms operates.

Quriobot

The first chatbot creation platform examined was Quriobot (https://quriobot.com) (accessed on 28 September 2022). It consists of an all-in-one solution for chatbot creation, as it provides the user with comprehensive options for creating and customizing the final product. This platform has been utilized in the past for journalistic purposes and, among other things, for implementations, including information collected from users [34]. The Dutch public broadcasting company KRO-NCRV is also listed as an official partner on the company website. The Quriobot platform operates by allowing users to string together several "steps" to create the final product (Figure 2). The end-user can navigate the conversation by clicking on pre-assigned buttons that guide them through the step sequences. Each step represents a specific action with predetermined ways of interaction between the user and the bot, so the complexity of the chatbot hinges on the variety of

steps available. More steps are being added over time, but as of the time of writing, a good selection already exists. Examples of steps include questions with a simple button for a Yes/No answer, open-ended questions that allow users to type out a response, contact forms, fields for uploading files, and so on. While there is a certain degree of choice for the user during their interaction–as with any chatbot that is properly thought out–Quriobot isn't capable of recognizing text input as a means of navigating the interface. As mentioned above, users can still input text in specific scenarios, but that text is only stored as an answer inside an internal repository for the journalist to decipher at a later time.

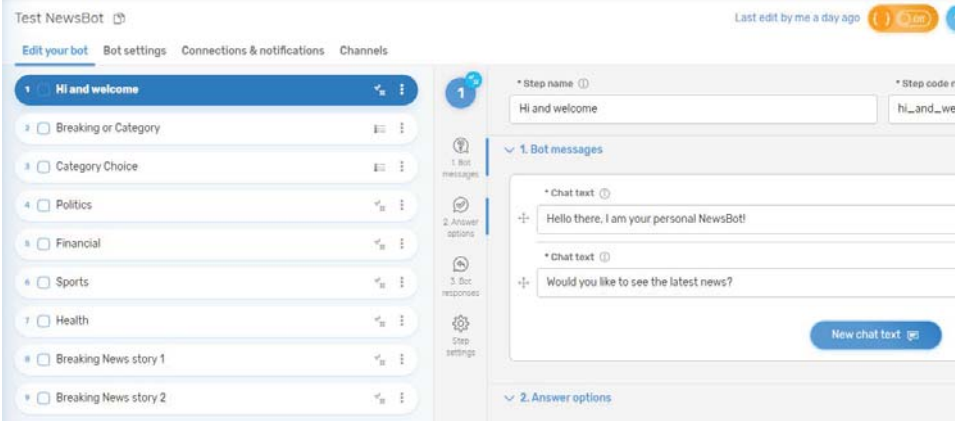


Figure 2. The chatbot creation interface of Quriobot. The "steps" can be seen on the left side.

SnatchBot

SnatchBot (https://snatchbot.me) (accessed on 28 September 2022) is the first creation tool on the list that exhibits natural language processing capabilities and does so in a coding-free manner, in addition to the more basic creation method that resembles the other two platforms (Figure 3). This tool is capable of understanding the user's intent, proper training, and acting accordingly to fulfill their requests. The way this is accomplished is by having the chatbot distinguish between two different types of text inputs, named Entities and Intent. Entities correspond to anything that can be named, such as objects, places, time, and so on. Intent, on the other hand, refers to the purpose behind the user's words in a sentence, like, for example: "I want to see the latest news". Aspiring chatbot creators can use this model by providing it with a training dataset for each of the two categories and then running the "train" command within the interface. The platform's processing capabilities will then take over to "teach" the chatbot to recognize certain things based on the provided data. If journalists are after the creation of a specialized chatbot, they can import—or create from scratch within the interface–the appropriate training data. However, SnatchBot also provides a handful of pre-trained models, even for free users, that can be utilized to recognize things like places, dates, currency-related terms, and even negative and positive words, which can prove very helpful for certain tasks like sentiment analysis. These models are actually recommended by the platform as the default setting since they cover a wide variety of situations.

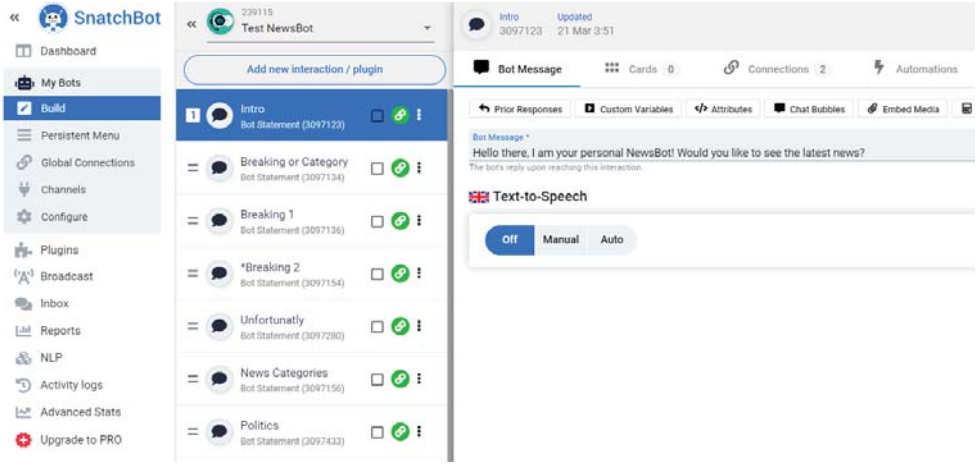


Figure 3. The chatbot creation interface of SnatchBot.

Chatfuel

Out of the three tools examined in this paper, Chatfuel (https://chatfuel.com) (accessed on 28 September 2022) is the platform with the most public association with the media industry. News agencies like MSNBC, BuzzFeed, and the Australian Broadcasting Corporation, to name a few, have all been listed in the past as official partners of the company on the Chatfuel website. On top of that, Forbes has also publicly stated their partnership with Chatfeul for the creation of their Telegram bot. This creation tool attempts to combine ease of use with a variety of advanced features. The way the platform operates is not dissimilar to Quriobot in the sense that all actions are represented by "blocks" (Figure 4). Each individual block can contain multiple actions, thus making them more feature-rich compared to the previously seen "steps". The creator can use visual programming to combine these blocks into a sequence that is usable by the audience. The most appealing characteristic of Chatfuel, however, is its ability to allow the audience to navigate the interface not only with buttons but with the use of natural language as well by typing out exactly what they want to do next, which allows them to bypass the predetermined path set by the journalist in favor of jumping directly to the task that is most relevant to them. The way this works from a creation standpoint is via the use of pattern matching, as the journalist can assign certain keywords to specific actions. The program will then try to identify those keywords within the user's text and forward them to the most relevant block. The keyword creation process can be as simple as the addition of a few shortcuts or as complex as a full list of recognized phrases, creating the illusion of an intelligent conversational agent. While this process is often very successful in convincing the end-user that the chatbot understands them, in reality, there is no natural language processing going on, but rather a comparison between their text input and a repository of responses in order to determine the most relevant answer, unlike the previously seen example of SnatchBot which is also capable of building "generative" chatbots, given the proper training.

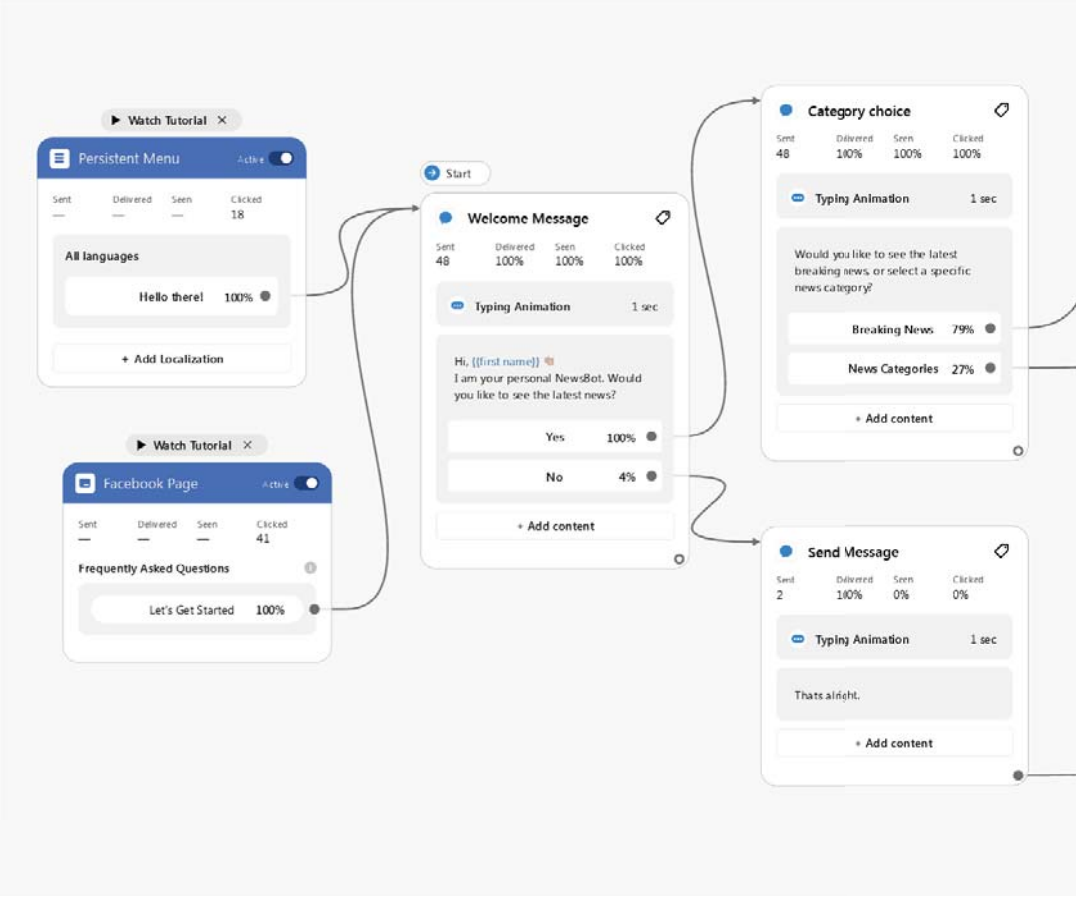


Figure 4. The chatbot creation interface of Chatfuel. "Blocks" are interconnected similarly to a flow diagram.

3.2.2. Heuristic Metric-Based Evaluation of Creation Interactivity

This manuscript aims to catalog the interactivity features of the selected platforms and facilitate a more in-depth look into their specific characteristics to distinguish between them. For that purpose Heeter's [49] model will be used, a theoretical construct that is considered by many to be one of the best attempts at categorizing interactivity, specifically in the realm of media and communication. Heeter's model has primarily been used for online websites, but it can be adapted very well for our purposes, as it allows us to quantify the interactivity of chatbot creation platforms based on the following six parameters:

1. The complexity of choice available: a metric of the provided features and the ability to customize certain parameters;
2. Effort users must exert: a measurement of how user-friendly the platform is, where the higher the score, the less effort is required for its proper use;
3. Responsiveness to the user: the extent of the ability of the end-user (audience) to contact the creator of the chatbot (journalist) through the interface;
4. Facilitation of interpersonal communication: the chatbot's ability to act as a medium through which users can communicate with each other;

5. Ease of adding information: a user's ability to contribute to the product by providing their own information and/or feedback;
6. Monitor system use: the ability of the creator to track and measure certain parameters regarding their chatbot, such as user statistics and behaviors.

These six dimensions allow for a very comprehensive and quantitative characterization of the available features found in online creation platforms. As the first step in the proposed three-way evaluation process, a rating between 1 and 5 on a Likert scale was assigned for each one of these categories by the authors. As for the specific labels of the scale, those will be explained in more detail in the next section. In addition to the above, other than rating the selected creation platforms using Heeter's model, each tool was also assigned to one of the interactivity categories described in Section 2.3, according to the model proposed by Jensen [47]. We utilized these two models in tandem for this step of the evaluation, as one of them serves as a more general categorization of all the examined tools, whereas the other provides more specific details when it comes to the individual aspects of each platform.

3.2.3. Chatbot Creation Workshop and Back-End Usability Evaluation

To evaluate the intuitiveness of the back-end of each platform (the programming of each chatbot) and the feasibility of the creation of chatbots by journalists with minimum background knowledge in computer science, a two-part workshop was organized. To begin with, as the second step in the evaluation process, a small group was formed, consisting of eight students from the School of Journalism of the Aristotle University of Thessaloniki. All participants were enrolled in the course of Human-Computer Interaction, an introductory course on application UX design and principles and usability evaluation. The participants had no previous experience with chatbot design. The workshop provided a short hands-on tutorial on every platform. The attendees were then requested to create a simple, retrieval-based chatbot for personalized news reporting. Participants spent 30 min on each platform, including the short tutorial and the time given to build the chatbot. In the end, they were asked to complete a survey concerning the usability of every platform and their experience with it. The workshop concluded with a short discussion, where participants had the opportunity to express their opinions and specific suggestions on their overall experience, as well as the different platforms.

After cataloging the experiences of individuals that are adjacent to the field of journalism, it was deemed necessary to also incorporate the opinions of working professionals in the field. To that end, the same methodology was used in order to organize a second workshop, where a small discussion group was formed consisting of seven professional journalists who underwent the exact same procedure described above. All participants were currently working–or have recently worked–in the field of journalism as of the time of the workshop and held no experience in regard to chatbot design. After the procedure was concluded, they were asked to share their opinions in regard to the creation platforms and their end products, in addition to filling out the aforementioned questionnaire.

3.2.4. Front-End Usability Evaluation

For the evaluation of the front end of each platform (interacting with each chatbot), an online survey was conducted as the final step of the proposed evaluation framework. In this experiment, the goal is to evaluate how the audience interacts with journalistic bots that have been created using the platforms under evaluation. Combined with the results from the workshop, this experiment is designed to provide insight concerning not only the usability, usefulness, and engagement of the process of chatbot programming but also how the result appeals to a broader audience. To address this, three simple chatbots were created by the research team, offering the same functionality for personalized news reporting (The sample chatbots that were used for the purposes of the survey can be found here: (1) https://botsrv2.com/qb/aPW6jrq9yvrR4ZXQ/eBYgZbjDd4r3l7jA?mode=preview (2) https://webbot.me/2eb10100d18b67dadeb5e732e5e9ff861aeed0639f80039e2c1e8c5d099b3bc8 (3) https://www.messenger.com/t/2253824708194900) (accessed on

30 May 2022). The respondents were asked to interact with each chatbot and then fill out a survey concerning the usability and their user experience with every chatbot. The survey was based on two of the most well-known frameworks for application evaluation, focusing on user interface satisfaction [50], as well as perceived usefulness and ease of use of the software [51], and a questionnaire was adapted for the specific needs of the conducted experiment. The survey was completed by a total of 162 participants, all students of the School of Journalism of the Aristotle University of Thessaloniki, enrolled in classes related to digital media.

During the survey preparation, all ethical approval procedures and rules suggested by the "Committee on Research Ethics and Conduct" of the Aristotle University of Thessaloniki were followed. The respective guidelines and information are available online at https://www.rc.auth.gr/ed/ (accessed on 8 June 2022). Moreover, the declaration of Helsinki and the MDPI directions for the case of pure observatory studies were also taken into account. Specifically, the formed questionnaire was fully anonymized, and the potential participants were informed that they agreed to the stated terms upon sending their final answers, while they had the option of quitting anytime without submitting any data.

4. Results and Discussion

4.1. Metric-Based Evaluation Results

Starting with the heuristic metric-based evaluation, the three chatbot creation platforms were judged depending on the features they offer. To begin with, each one of the selected tools was cataloged according to Jensen's categorization [47], as explainedin previous chapters. After this process, the three chatbot creation platforms were assessed by the authors. Each one of the four judges spent the same amount of time with each of the platforms in question and then proceeded to rate the features offered by ranking them according to the six interactivity parameters mentioned in Heeter's model [49], evaluating them using a Likert scale from 1 to 5. Clear guidelines were established for the rating procedure and for what each one of the scores on the scale represents. Specifically: A score of 1 denotes the complete absence of a feature. A score of 2 represents the existence of a feature that has a very barebones implementation (for example, very limited user-to-user interactions). A 3 is used as a baseline and is assigned to a feature that is decently developed, but within this standardized categorization is stillmissing some features compared to the competition. A score of 4 indicates a very well-implemented feature that addresses a subject from different angles (like the inclusion of multiple different ways for the audience to contact the creator of the bot). Finally, a 5 indicates a highly advanced implementation for that specific characteristic that includes a multitude of features compared to the competition (for example, a very robust analytics system that records detailed interactions between user and bot). To ensure inter-rater reliability between the four judges, a percent agreement system was used [52], and the formation of the subsequent result matrix revealed a very high agreement between the raters, with a very small amount of outlying scores. Specifically, following this clearly outlined system, the four raters proposed the exact same score for all categories, with the only exception being the SnatchBot platform, in regards to the category "Effort users must exert". This category was rated differently between the four reviewers and received an average score of 3, as there was some slight degree of fluctuation in the opinions of the raters in regard to the ease of use of some of the features provided by this tool. This outlying observation could be the result of the highly modular system presented by SnatchBot, which left some room for interpretation as to the degree of competency required by the average user to create a functional bot. An overview of the final verdict of the interactivity features comparison, which consists of the mean values of all four raters, can be seen in Table 1 below.

Table 1. Chatbot creation platforms and their interactivity features ranked.

Interactivity Parameters	Quriobot	SnatchBot	Chatfuel
Complexity of choice available	3	5	5
Effort users must exert	4	3	4
Responsiveness to the user	4	4	5
Facilitation of interpersonal communication	1	2	2
Ease of adding information	4	4	4
Monitor system use	5	2	3
Total:	21	20	23

The Quriobot platform was the first one to be examined, and based on the implementation of the features described in the previous chapter, it was assigned to the category of Consultational interactivity tools. Users are able to request specific things from a chatbot created through this platform. However, that can't be accomplished through text input but rather through scripted interactions. As for its individual scores based on the Heeter model, they are as follows: Complexity of choice available: 3, Effort users must exert: 4, Responsiveness to the user: 4, Facilitation of interpersonal communication: 1, Ease of adding information: 4, Monitor system use: 5, for a total of 21.

The available features Quriobot offers during the creation process can be characterized as adequate compared to other similar tools. The standout characteristic of this platform, however, is its extensive monitoring capabilities, which are part of the reason why this tool excels in the creation of chatbots aimed at data acquisition from users. All information collected by the bot during its interactions is saved and can be accessed by the journalist in an excel sheet-style interface. This includes direct user input like text and uploaded files, the exact "Step Path" that a user followed when navigating the interface, as well as comprehensive metadata like time of access, the user's operating system, and even location. On the other end of the spectrum, Quriobot doesn't facilitate user-to-user interaction in any way. Thus the lowest possible score was assigned for the interpersonal communication category. Specific steps that present contact forms, text input, and the ability for users to rate the chatbot also exist, so the other categories also receive an above-average score. Overall the ease of use for this platform is quite high, as it allows for visual programming, which can ease the uninitiated into chatbot creation for the first time.

The capabilities exhibited by SnatchBot make it the only one of the examined tools that belong in the category of Registrational interactivity, as it can potentially understand the users and the intent behind their queries. Of course, this doesn't mean that all products created with this platform will necessarily belong in this category, as it is up to the user as to how much they want to invest in the training of the NLP models, if at all, as SnatchBot also provides all the necessary tools to create a non-intelligent agent as well. As for its scores based on the Heeter model: Complexity of choice available: 5, Effort users must exert: 3, Responsiveness to the user: 4, Facilitation of interpersonal communication: 2, Ease of adding information: 4, Monitor system use: 2, for a total of 20.

These natural language processing capabilities open up a whole realm of possibilities for the journalist, and even though their full scope of applications exceeds the narrower focus of this paper, SnatchBot still presents a compelling alternative for those who require a more engaging user experience out of their bot. The ease of use does suffer slightly since the whole process of training an NLP model isn't quite as straightforward as the simple visual programming of other platforms, but user responsiveness remains high because of the combination of NLP and the existence of a specific plug-in to pass the conversation onto a human. Multiple fields for information input exist–although there is no specific rating feature–and the monitoring statistics offered to free users are fairly limited, which is reflected in the final score of the above categories. In the past, integration with Chatbase–a popular chatbot analytics service provided by Google–used to be present,

but unfortunately, this doesn't mitigate the monitoring shortcomings of the platform, as this service was officially discontinued in 2021 after operating in maintenance mode for an extended period, leaving SnatchBot without any noteworthy monitoring capabilities to speak of.

Lastly, the options present in the third and final platform, Chatfuel, firmly place it in the realm of Conversational interactivity. This is due to its ability to interact with users both with predetermined buttons as well as free text, but without presenting options that can result in a truly generative dialogue between the bot and a human. Going into specifics, Chatfuel received the scores below: Complexity of choice available: 5, Effort users must exert: 4, Responsiveness to the user: 5, Facilitation of interpersonal communication: 2, Ease of adding information: 4, Monitor system use: 3, for a total of 23.

Complexity of choice and user responsiveness are the strongest suits of this platform. The sheer amount of options that Chatfuel's pattern-matching solution offers will likely be enough to satisfy all but the most demanding use cases in the realm of journalistic work. The keyword system essentially makes the final product proportionally intelligent to the amount of work the user is willing to invest, making it versatile but not overwhelming to newcomers, thus the high score in the first two categories. Additionally, Chatfuel provides high-quality features in the "Responsiveness to the user" category, as it is the only platform capable of accommodating a live chat with the bot administrator, in addition to the typical conversation handover feature that can pass the text chat on to a human. Some very limited user-to-user communication features are present, but as expected from most chatbots, this isn't the main feature of the platform. Variables can also be used to extract some user information (for example, a name to address them by) in the case that users engage with the bot via a registered account, like, for example, their Facebook profile. Finally, while there is analytics available, a significant number of features are inaccessible to free users, as they require a subscription, and so the "Monitor system use" category receives a middle-of-the-road score.

What can be surmised from the categorization based on Heeter's [49] model is that, despite an abundance of features in most cases, the particular interactivity characteristic where chatbot creation platforms underperform is the facilitation of interpersonal communication. This is understandable to an extent, as, in most instances, the use-case for conversational agents dictates the need for back-and-forth communication between the chatbot and the audience–or in some cases, the audience and the creator of the program—but it doesn't account for the user-to-user department. The market could attempt to adapt to the lack of features in this area, as user-to-user interaction via chatbots could facilitate the easier creation of online communities based around media organizations or even the enhancement of already existing communities through more direct and interactive connections among members.

Most of the examined platforms can satisfy the interactive needs of journalists in a variety of scenarios, as user responsiveness and ease of adding information remain high across all platforms. Another thing that could be improved, however, is the monitor system use category, as some inconsistencies can be observed over different platforms, with some of them offering extensive features and others only supplying the bare minimum for the bot administrator. This can be specifically observed in platforms that create a large separation between their free and premium plans, as analytics seems to be presented as one of the prime incentives for users to upgrade.

Overall, a decent spread of features seems to be present in the current market, as options exist even for more demanding users that require natural language processing (SnatchBot) as well as users looking for more basic–but still robust–interactions, with emphasis on information gathering (Quriobot). Chatfuel specifically represents an example of a platform that can potentially appeal to an even larger part of the media sphere because of its more generalist approach and the complexity of choices it offers. The existence and variety of these platforms, as well as their accessibility, suggest a future where news organizations can lean even more into their interactive side and create a news-sharing envi-

ronment that will provide users with a level of back-and-forth communication previously unseen across the media industry.

4.2. Back-End Workshop and Evaluation Results

As described in Section 3.2.3, a two-part workshop was used as an evaluation method for each of the chatbot creation interfaces. This endeavor aimed to directly compare and contrast the usability of all the features present in each of the platforms and to examine how each one of them could realistically be utilized both by professional journalists as well as students in the field of media. A hands-on approach was adopted, where each participant was asked to create their own chatbot by the end of the procedure. The first part of the workshop, pertaining to the opinions of the students, took place during the first days of March 2022 and was approximately two hours long. The second part, which studied the perceptions of professional journalists, took place at the beginning of September 2022 and had a similar length.

For the purposes of this workshop procedure, it was important to procure participants that were all adjacent to the field of journalism but with varying degrees of familiarity with the profession. When it comes to the first part of this two-part workshop, the participants (n = 8, 4 male and 4 female) were between the ages of 18 and 30, with no prior experience or specialization in the field of ICT or chatbot creation. They were all students in the School of Journalism of the Aristotle University of Thessaloniki, enrolled in the course on Human-Computer Interaction, thus making them qualified to accurately judge the design process through a journalistic approach. As for the second part of the workshop, the participants (n = 7, 4 male and 3 female) were all professional journalists with working experience in the field but without any specialization in ICT or prior knowledge when it comes to chatbot creation. Their age group ranged from 31 to 60 years old.

After a detailed presentation of the creation process for each one of the platforms, the attendees were asked to allocate some time to create a sample chatbot for each one of the selected tools. Instructions were given for the specifications of the sample chatbot, but no extra help was provided to them to more accurately gauge their level of understanding for each of the creation interfaces. To ensure the validity of the results, the subjects were not allowed to communicate with each other during this stage of the study. After this process was concluded, participants were asked to fill out a questionnaire pertaining to their experience. The evaluation survey results are shown in Figure 5, and they are measured on a seven-point Likert scale where 1 represents a negative connotation, and 7 represents a positive one.

As can be seen from the above, the results of the workshop show that all three platforms were generally well-received. Quriobot, however, was consistently ranked higher than its peers in all categories. This lead ranges from small to relatively significant, and even though it spans all metrics, it seems to stand out, particularlyin terms of ease of use and straightforwardness. Specifically, Quriobot received an average rating of 5.91 (SD = 1) in the ease-of-use category, compared to the other two platforms, which ranged from 4.25 (SD = 1.3) for SnatchBot to 4.58 (SD = 1.5) for Chatfuel. Similarly, the process of correcting user mistakes was deemed to be noticeably easier on Quriobot with a mean score of 5.75 (SD = 0.8) as opposed to SnatchBot (M = 4, SD = 1.3) and Chatfuel (M = 4.33 SD = 1.6). Statistics such as these swayed user opinions into ranking the chatbot creation interface of Quriobot higher than the other two alternatives, as participants concluded that it is more suited for all levels of users (M = 4.75, SD = 1.2) compared to the alternatives (SnatchBot: M = 2.91, SD = 0.9; Chatfuel: M = 3.41, SD = 1.5).

Figure 5. Box Plot of evaluation results from the two-part chatbot programming workshop.

When it comes to specific comments left by the participants during the discussion portion of the workshop, the declared opinions seem to substantiate the recorded results. During the first part of the workshop, Quriobot was deemed the most accessible platform overall, with one participant proclaiming that "it had the most friendly and self-explanatory environment". Despite that, however, another participant did point out a visual bug they noticed in the interface of the program. Chatfuel had a similar, albeit less enthusiastic, reception, with attendees noting that it sported a "beautiful design" with "easy to connect flows" but also noticing that "it was slightly harder to initiate the process of chatbot creation on it". Lastly, SnatchBot was the interface that received the most critical comments overall, with participants suggesting that "it seems only suitable for pro users" and saying that "it was difficult to understand my mistakes in this platform". Despite that, however, it was still positively received overall, as can be seen by the aggregate results shown earlier, with two users pointing out that it presented them with the greatest variety of options, themes and topics.

When it comes to the second part of the workshop procedure, the professionals were mostly surprised by the simplicity and accessibility of modern approaches to chatbot design. All three platforms had a positive reception overall, with Quriobot standing out as slightly easier to use compared to the other two. One participant who had a disappointing experience trying to program a chatbot in the past and abandoned the endeavor very early now felt confident using the three presented tools. Another one declared that they had no experience, but they felt confident in engaging with the presented platforms regardless, without having any expertise. The same participant stated that even though they "appreciated the intuitive design of Quriobot", they would like to explore the other two platforms more to make use of what they perceived as "their extended functionality". Concerning the integration in journalistic workflows, one attendee found the possibility of constantly updating personalization parameters through a chatbot a

particularly interesting use case, especially for smaller news organizations. Some concerns were also voiced by a workshop participant in regard to how the chatbot could operate with open-ended questions in a more conversational environment. Furthermore, the same professional noted that they would like to see evaluation results concerning their long-term use to see proof that parametrization and personalization will, in fact, enhance engagement over time. One concern that also arose during this second part of the workshop was that, according to a participant's experience, modern users tend to mostly browse article titles and short descriptions, and they might not dedicate extra time in conversation with a chatbot for a deeper exploration of a news agenda.

Overall, besides the usability metrics, all three platforms were also perceived as useful for journalistic purposes, which is particularly relevant for the purposes of this study, as it adheres to the notion that these platforms can be utilized by the average journalist for their work. This notion extended to both university students, as well as professional journalists. Once again, Quriobot seems to be a standout in terms of perceived usefulness. This lead, however, is smaller compared to other categories and given the limited sample size of the workshops, as well as the possibility of utilizing each platform for different purposes in the real world, it is safe to assume that all three of the alternatives present a compelling case for being implemented into modern journalistic practice.

4.3. Front-End Evaluation Results

A total of 162 participants (49 male, 113 female) responded to the questionnaire, which was disseminated via the LimeSurvey (https://www.limesurvey.org) (accessed on 30 May 2022) platform during April and May of 2022. Ninety-two percent of them were in the age group of 18–24, 2% were between 25–40 and 6% were in the category of 40–60. Fifty-seven percent of them held a high school diploma, and 43% had a university or postgraduate degree. Seventy-seven percent declared that they have more than average experience with IT and communication technologies, and 81% of them have interacted in the past with a chatbot or another automated application, using a personal computer (48%), a mobile phone (44%) or other devices (2%). The results of the online questionnaires are shown in Figure 6.

Results show that users were, on average more eager to communicate with a chatbot created by Quriobot or Snatchbot, as they found the environment of Chatfuel unnecessarily complex and less easy to use compared to the other two. This increases the likelihood that a user might require assistance from a technical expert to use this tool, which seems rather inconvenient, given the examined use-case scenario. Participants believed that most people would quickly learn to use Quriobot and Snatchbot, while they felt that they would need to learn more things before using Chatfuel. Functions seemed to be overall better integrated into Quriobot (M = 3.83, SD = 0.80), as this platform was considered quite consistent, in contrast to Chatfuel (M = 3.62, SD = 0.79) and Snatchbot (M = 3.18, SD = 1.08) which exhibited slightly lower scores in this category. Overall, respondents felt more confident using Quriobot as an automated personalized newsagent.

It is worth mentioning that these results refer to users interacting with the chatbots for approximately 1–2 min each, so they represent a first contact scenario with the platforms. This means that results may not be indicative of the long-term user experience, but rather they more accurately represent the learning curve of each platform. In addition, despite the gradation of the results and the subsequent highlighting of the user's preference for the Quriobot agent, all three platforms were considered easy to use, and participants generally felt confident using them after only small testing. This fact is particularly encouraging for the feasibility of using automated chatbots for personalized news reporting, as it suggests that journalists without a technical background, prior training, or any further instructions can immediately feel confident and engaged using them. This is part of the gap that this paper aims to fill in the related literature, and it proves to have a significant connotation when observed in conjunction with the evaluation results of the frontend of each application.

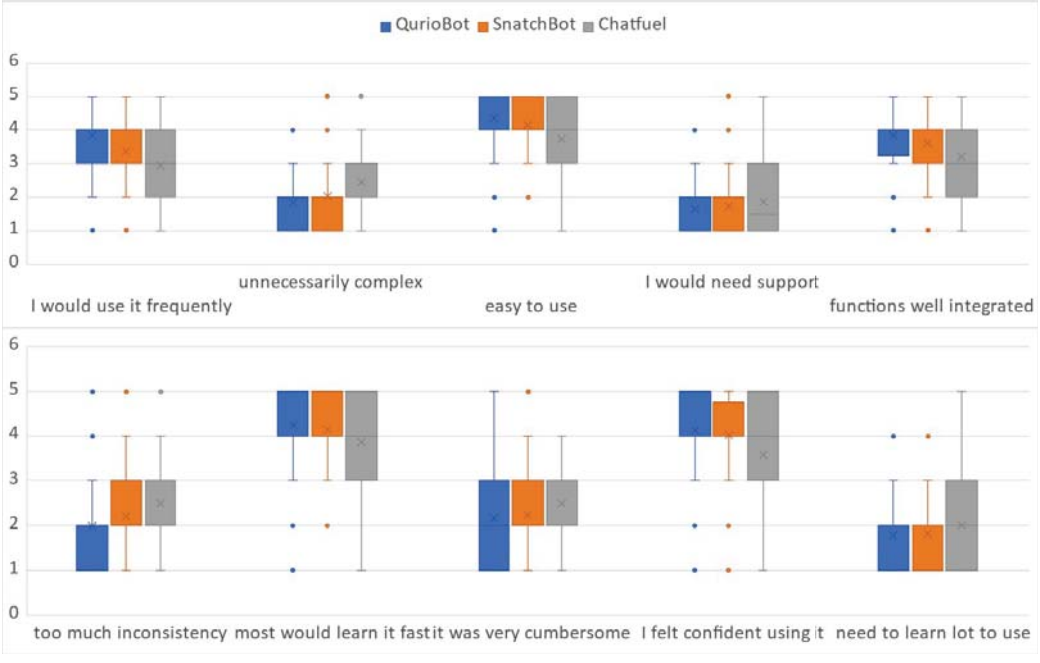


Figure 6. Box Plot of evaluation results from the chatbot user experience.

5. Conclusions and Further Work

The operational principles of conversational agents render them interactive pieces of software, almost regardless of the context they are being used. Those intrinsic characteristics make them an invaluable ally to the modern journalist, as they can be used to enhance particular aspects of their workflow that could benefit from systematic automation and interactivity, like, for example, news dissemination. To that end, this study focused on a multifaceted evaluation approach for existing chatbot creation platforms, bringing the average journalist to the forefront and examining the feasibility of integrating chatbot automation into day-to-day journalistic practices.

Chatbot creation nowadays is not as inaccessible as it once was, as even the average worker in the media industry can pick up and use one of the many available online tools that are based on visual scripting, and the market is saturated with a variety of different options to cover the needs of different users. As suggested by the workshop that was carried out during this study, the average journalist today is likely to be capable of creating a basic chatbot by utilizing the available online tools, with minimum to no guidance, even without previous experience. Furthermore, what the results of the study uncovered were that the proposed three-way evaluation approach can potentially be used to provide a more well-rounded view of existing tools that includes an examination of both the front-end as well as the back-end of an application, in addition to specific interactivity features that are considered essential for use in a media environment. This spherical examination can lead to the formation of a more balanced opinion when it comes to selected tools and platforms, one that can highlight potential inconsistencies between the performance of a specific tool when it comes to its creation process as opposed to the process of using it. For instance, out of the curated list of online tools chosen for this study, Quriobot was overall the most well-received one, both in terms of usability, as well as usefulness. This preference extends to both the back-end, as well as from the front-end of the application, as it was ranked the highest both during the two-part workshop, as well as the end-user survey. However, some inconsistencies can be observed between the scores of the other two platforms, depending

on the evaluation scenario. For example, workshop attendees characterized SnatchBot's interface as relatively hard to use, noting that it caters to more professional users. This was especially notable on behalf of the university students during the first part of the workshop and less so during the second part, which included professional journalists, although it was still prevalent. On the other hand, however, this was not true for the end-users, who evaluated the end product of SnatchBot as relatively easy to use. This dissonance highlights an important point for this study, which is the fact that when it comes to chatbot usage in a journalistic environment, all aspects of a potential tool should be examined separately, as usability and usefulness are both multifaceted terms that affect the journalist and the end-user in different ways, due to the unique nature of conversational agents.

Overall, all three platforms received relatively high scores for usefulness and usability, making them all suitable alternatives for aspiring chatbot creators in the media industry. Concerning the opinions of professional journalists specifically, the overall conclusion was that participants felt confident in using all three platforms for chatbot design. This is a noteworthy result, considering that before the short presentation of the three platforms, the attendees were either unaware or skeptical when it came to the easy integration of such tools in their work. Having said that, there are a few particular areas where improvements could still be made to the available tools in the market to ensure they cater to all types of interactivity and refine the already available features. As an example–stated in the corresponding section–some interactivity features, like the facilitation of interpersonal communication, are not yet integrated into many of the available platforms. Their incorporation could help improve those tools and saturate the market with more niche options for chatbot creation in the field of journalism and media. Furthermore, professionals raised a few concerns in regard to chatbot usage and how it relates to user engagement, which is an interesting point that could be addressed in future research over a real-world case study. Taking this into account, this work could also be elaborated on in the future with the inclusion of different evaluation techniques that focus on hyper-specific aspects of current journalistic practices and figuring out which available platforms better suit each task at hand. For example, a potential study could include several platforms, which would all be evaluated solely on the basis of how they perform when it comes to information gathering from audience members or other similarly distinct tasks.

Author Contributions: Conceptualization, E.K. and A.V.; methodology, E.K. and N.V.; validation, A.V. and C.D.; formal analysis, E.K. and N.V.; investigation, E.K., N.V., C.D. and A.V.; resources, N.V., C.D. and A.V.; data curation, E.K. and N.V.; writing—original draft preparation, E.K., N.V., C.D. and A.V.; writing—review and editing, E.K., N.V., C.D. and A.V.; visualization, E.K., N.V. and C.D.; supervision, C.D. and A.V.; project administration, C.D. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to consent provided by the participants, being informed prior to their involvement in the evaluation. Furthermore, procedures and rules suggested in the reference handbook of the "Committee on Research Ethics and Conduct" of the Aristotle University of Thessaloniki were fully adopted.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spyridou, L.-P.; Matsiola, M.; Veglis, A.; Kalliris, G.; Dimoulas, C. Journalism in a State of Flux: Journalists as Agents of Technology Innovation and Emerging News Practices. *Int. Commun. Gaz.* **2013**, *75*, 76–98. [CrossRef]
2. Pavlik, J. The Impact of Technology on Journalism. *Journal. Stud.* **2000**, *1*, 229–237. [CrossRef]
3. Kotenidis, E.; Veglis, A. Algorithmic Journalism—Current Applications and Future Perspectives. *Journal. Media* **2021**, *2*, 244–257. [CrossRef]

4. Valtolina, S.; Barricelli, B.R. Chatbots and Conversational Interfaces: Three Domains of Use. In Proceedings of the Fifth International Workshop on Cultures of Participation in the Digital Age: Design Trade-offs for an Inclusive Society co-located with the International Conference on Advanced Visual Interfaces, Castiglione della Pescaia, Italy, 29 May 2018; pp. 62–70.
5. Rogers, E.M. *Diffusion of Innovations*, 3rd ed.; Free Press: New York, NY, USA, 1983; ISBN 978-0-02-926650-2.
6. Chung, D.S. Profits and Perils: Online News Producers' Perceptions of Interactivity and Uses of Interactive Features. *Convergence* **2007**, *13*, 43–61. [CrossRef]
7. Jain, M.; Kumar, P.; Kota, R.; Patel, S.N. Evaluating and Informing the Design of Chatbots. In Proceedings of the 2018 Designing Interactive Systems Conference, Hong Kong, China, 8 June 2018; pp. 895–906.
8. Dale, R. The Return of the Chatbots. *Nat. Lang. Eng.* **2016**, *22*, 811–817. [CrossRef]
9. Ritter, A.; Cherry, C.; Dolan, W.B. Data-Driven Response Generation in Social Media. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 583–593.
10. Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; Li, Z. Sequential Matching Network: A New Architecture for Multi-Turn Response Selection in Retrieval-Based Chatbots. *arXiv* **2017**, arXiv:1612.01627.
11. Veglis, A.; Maniou, T.A. Chatbots on the Rise: A New Narrative in Journalism. *Stud. Media Commun.* **2019**, *7*, 1–6. [CrossRef]
12. Li, X.; Mou, L.; Yan, R.; Zhang, M. StalemateBreaker: A Proactive Content-Introducing Approach to Automatic Human-Computer Conversation. *arXiv* **2016**, arXiv:1604.04358.
13. Shang, L.; Lu, Z.; Li, H. Neural Responding Machine for Short-Text Conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; Association for Computational Linguistics, Beijing, China, 26–31 July 2015; Volume 1, pp. 1577–1586.
14. Kim, J.; Lee, H.-G.; Kim, H.; Lee, Y.; Kim, Y.-G. Two-Step Training and Mixed Encoding-Decoding for Implementing a Generative Chatbot with a Small Dialogue Corpus. In Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG), Tilburg, The Netherlands, 5 November 2018; pp. 31–35.
15. Mondal, A.; Dey, M.; Das, D.; Nagpal, S.; Garda, K. Chatbot: An Automated Conversation System for the Educational Domain. In Proceedings of the 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 15–17 November 2018; pp. 1–5.
16. Janarthanam, S. *Hands-on Chatbots and Conversational UI Development: Build Chatbots and Voice User Interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*; Packt Publishing Ltd.: Birmingham, UK, 2017.
17. Weizenbaum, J. ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
18. Shum, H.-Y.; He, X.; Li, D. From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. *arXiv* **2018**, arXiv:1801.01957. *19*, 10–26. [CrossRef]
19. Anderson, C.W. Towards a Sociology of Computational and Algorithmic Journalism. *New Media Soc.* **2013**, *15*, 1005–1021. [CrossRef]
20. Salzwedel, M. The Rise of Robojournalism: African Trends. *Rhodes Journal. Rev.* **2014**, *2014*, 85–87.
21. Carlson, M. Automating Judgment? Algorithmic Judgment, News Knowledge, and Journalistic Professionalism. *New Media Soc.* **2018**, *20*, 1755–1772. [CrossRef]
22. Lokot, T.; Diakopoulos, N. News Bots: Automating News and Information Dissemination on Twitter. *Digit. Journal.* **2016**, *4*, 682–699. [CrossRef]
23. Jones, B.; Jones, R. Public Service Chatbots: Automating Conversation with BBC News. *Digit. Journal.* **2019**, *7*, 1032–1053. [CrossRef]
24. Diakopoulos, N. *Automating the News*; Harvard University Press: Cambridge, MA, USA, 2019.
25. Veglis, A.; Dimoulas, C.; Kalliris, G. Towards Intelligent Cross-Media Publishing: Media Practices and Technology Convergence Perspectives. In *Media Convergence Handbook—Vol. 1: Journalism, Broadcasting, and Social Media Aspects of Convergence*; Lugmayr, A., Dal Zotto, C., Eds.; Media Business and Innovation; Springer: Berlin/Heidelberg, Germany, 2016; pp. 131–150. ISBN 978-3-642-54484-2.
26. Thurman, N. Computational Journalism. In *The Handbook of Journalism Studies*; Wahl-Jorgensen, K., Hanitzsch, T., Eds.; Routledge: New York, NY, USA, 2018; pp. 180–195. ISBN 9781315167497.
27. Graefe, A. Guide to Automated Journalism. Available online: https://academiccommons.columbia.edu/doi/10.7916/D80G3XDJ (accessed on 28 September 2022).
28. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2018; ISBN 978-1-5225-5593-3.
29. Sidiropoulos, E.; Vryzas, N.; Vrysis, L.; Avraam, E.; Dimoulas, C. Growing Media Skills and Know-How In Situ: Technology-Enhanced Practices and Collaborative Support in Mobile News-Reporting. *Educ. Sci.* **2019**, *9*, 173. [CrossRef]
30. Clerwall, C. Enter the Robot Journalist: Users' Perceptions of Automated Content. *Journal. Pract.* **2014**, *8*, 519–531. [CrossRef]
31. Carlson, M. The Robotic Reporter. *Digit. Journal.* **2015**, *3*, 416–431. [CrossRef]

32. Matsiola, M.; Dimoulas, C.; Kalliris, G.; Veglis, A.A. Augmenting User Interaction Experience through Embedded Multimodal Media Agents in Social Networks. Available online: https://www.igi-global.com/chapter/augmenting-user-interaction-experience-through-embedded-multimodal-media-agents-in-social-networks/www.igi-global.com/chapter/augmenting-user-interaction-experience-through-embedded-multimodal-media-agents-in-social-networks/198632 (accessed on 12 June 2022).
33. Veglis, A.; Maniou, T.A. Embedding a Chatbot in a News Article: Design and Implementation. In Proceedings of the 23rd Pan-Hellenic Conference on Informatics, Nicosia, Cyprus, 28–30 November 2019; pp. 169–172.
34. Veglis, A.; Kotenidis, E. Employing Chatbots for Data Collection in Participatory Journalism and Crisis Situations. *J. Appl. Journal. Media Stud.* **2020**, *11*, 309–332. [CrossRef]
35. Maniou, T.A.; Veglis, A. Employing a Chatbot for News Dissemination during Crisis: Design, Implementation and Evaluation. *Future Internet* **2020**, *12*, 109. [CrossRef]
36. Verma, P.; Saxena, A.; Sharma, A.; Thies, B.; Mehta, D. A WhatsApp Bot for Citizen Journalism in Rural India. In *ACM SIGCAS Conference on Computing and Sustainable Societies*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 423–427.
37. Ford, H.; Hutchinson, J. Newsbots That Mediate Journalist and Audience Relationships. *Digit. Journal.* **2019**, *7*, 1013–1031. [CrossRef]
38. Shin, D.; Al-Imamy, S.; Hwang, Y. Cross-Cultural Differences in Information Processing of Chatbot Journalism: Chatbot News Service as a Cultural Artifact. *Cross Cult. Strateg. Manag.* **2022**, *29*, 618–638. [CrossRef]
39. Steuer, J. Defining Virtual Reality: Dimensions Determining Telepresence. *J. Commun.* **1992**, *42*, 73–93. [CrossRef]
40. Liu, Y.; Shrum, L.J. What Is Interactivity and Is It Always Such a Good Thing? Implications of Definition, Person, and Situation for the Influence of Interactivity on Advertising Effectiveness. *J. Advert.* **2002**, *31*, 53–64. [CrossRef]
41. Ksiazek, T.B.; Peer, L.; Lessard, K. User Engagement with Online News: Conceptualizing Interactivity and Exploring the Relationship between Online News Videos and User Comments. *New Media Soc.* **2016**, *18*, 502–520. [CrossRef]
42. Boczkowski, P.J. *Digitizing the News: Innovation in Online Newspapers*; MIT Press: Cambridge, MA, USA, 2005.
43. Deuze, M. What Is Journalism? Professional Identity and Ideology of Journalists Reconsidered. *Journalism* **2005**, *6*, 442–464. [CrossRef]
44. Chung, D.S.; Yoo, C.Y. Audience Motivations for Using Interactive Features: Distinguishing Use of Different Types of Interactivity on an Online Newspaper. *Mass Commun. Soc.* **2008**, *11*, 375–397. [CrossRef]
45. Veenstra, M.; Wouters, N.; Kanis, M.; Brandenburg, S.; teRaa, K.; Wigger, B.; Moere, A.V. Should Public Displays Be Interactive? Evaluating the Impact of Interactivity on Audience Engagement. In Proceedings of the 4th International Symposium on Pervasive Displays, Saarbrücken, Germany, 10 June 2015; pp. 15–21.
46. Brissette-Gendron, R.; Léger, P.-M.; Courtemanche, F.; Chen, S.L.; Ouhnana, M.; Sénécal, S. The Response to Impactful Interactivity on Spectators' Engagement in a Digital Game. *Multimodal Technol. Interact.* **2020**, *4*, 89. [CrossRef]
47. Jensen, J.F. 'Interactivity': Tracking a New Concept in Media and Communication Studies. *Nord. Rev.* **1998**, *19*, 185–204.
48. Asbjørn, F.; Petter, B.B.; Tom, F.; Effie, L.L.; Manfred, T.; Ewa, A.L. Chatbots for Social Good. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2018.
49. Heeter, C. Implications of Interactivity for Communication Research. In *Media Use in the Information Age: Emerging Patterns of Adoption and Consumer Use*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1989; pp. 217–235.
50. Chin, J.P.; Diehl, V.A.; Norman, K.L. Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 1 May 1988; pp. 213–218.
51. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
52. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Future Internet Editorial Office
E-mail: futureinternet@mdpi.com
www.mdpi.com/journal/futureinternet

Academic Open
Access Publishing

www.mdpi.com

ISBN 978-3-0365-7651-0