future internet

MDPI

Article

Methodological Approach for Identifying Websites with Infringing Content via Text Transformers and Dense Neural Networks

Aldo Hernandez-Suarez *, Gabriel Sanchez-Perez, Linda Karina Toscano-Medina, Hector Manuel Perez-Meana, Jose Portillo-Portillo and Jesus Olivares-Mercado

Instituto Politecnico Nacional, ESIME Culhuacan, Mexico City 04440, Mexico; gasanchezp@ipn.mx (G.S.-P.); ltoscano@ipn.mx (L.K.T.-M.); hmperezm@ipn.mx (H.M.P.-M.); jportillop@ipn.mx (J.P.-P.); jolivares@ipn.mx (J.O.-M.)

* Correspondence: alhernandezsu@ipn.mx

Abstract: The rapid evolution of the Internet of Everything (IoE) has significantly enhanced global connectivity and multimedia content sharing, simultaneously escalating the unauthorized distribution of multimedia content, posing risks to intellectual property rights. In 2022 alone, about 130 billion accesses to potentially non-compliant websites were recorded, underscoring the challenges for industries reliant on copyright-protected assets. Amidst prevailing uncertainties and the need for technical and AI-integrated solutions, this study introduces two pivotal contributions. First, it establishes a novel taxonomy aimed at safeguarding and identifying IoE-based content infringements. Second, it proposes an innovative architecture combining IoE components with automated sensors to compile a dataset reflective of potential copyright breaches. This dataset is analyzed using a Bidirectional Encoder Representations from Transformers-based advanced Natural Language Processing (NLP) algorithm, further fine-tuned by a dense neural network (DNN), achieving a remarkable 98.71% accuracy in pinpointing websites that violate copyright.

Keywords: dense neural network; privacy violations; illegal download; BERT; natural language processing; infringing content

Citation: Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, L.K.; Perez-Meana, H.M.; Portillo-Portillo, J.; Olivares-Mercado, J. Methodological Approach for Identifying Websites with Infringing Content via Text Transformers and Dense Neural Networks. *Future Internet* **2023**, *15*, 397. https://doi.org/10.3390/fi15120397

Academic Editor: Michael Sheng

Received: 8 November 2023
Revised: 6 December 2023
Accepted: 6 December 2023
Published: 9 December 2023

1. Introduction

The Internet, a complex network facilitating the exchange of information via websites and APIs, has greatly influenced various sectors like email, education, healthcare, and entertainment, optimizing data dissemination [1–3]. This evolution gave rise to the Internet of Everything (IoE), an architectural framework that enables access to a vast array of digital content [4]. Key areas of multimedia consumption include education, cinema, music, literature, gaming, streaming services, and software [5].

Legal measures, as per the Digital Millennium Copyright Act (DMCA), are essential to protect copyrights on IoE platforms against illegal copying and distribution, also referred to as infringing content [6,7]. However, The IoE's extensive volume poses challenges in monitoring and controlling the distribution of infringing content, despite legal and technical efforts to conduct takedowns [8,9]. Historical instances, like the legal actions against Napster in 2001 for pirated music and Megaupload in 2012 for unauthorized downloads, highlight the ongoing struggle against digital unauthorized sharing and its legal consequences [10,11].

Despite efforts to combat infringing content, the dissemination of IoE continues to grow. According to [12], the increased use of information technologies (IT) during the SARS-CoV-19 pandemic led to a significant rise in intellectual property theft, with an estimated 8.8 trillion infringed files dispersed by the end of 2021 through websites, video-streaming apps, online social media platforms (OSNs), and instant messaging channels.

Data from 2022 [13] confirm this trend, with search engines like Google Search [14] generating 68.6 million URLs linked to potentially bypassed authorship-claimed digital works.

The UK Government's Intellectual Property Office [15,16] annually issues the Online Copyright Infringement Tracking Report, which examines the browsing behavior of a representative sample of users based on their queries in various search engines. This analysis contributes to the classification of websites, guided by the nature and reputation of the content they offer, meticulously evaluating potential infringements of intellectual property rights. Within this framework, the categorization of websites in terms of copyright infringement is segmented into three categories: legal, mixed legal and infringing, and infringing sites. The first category, pertaining to legal sites, is dedicated to promoting, offering, and distributing content endorsed by the authors or their representatives, as well as works belonging to the public domain. The second category, the mixed ones, partially conform to intellectual property regulations, hosting both authorized works and those that are not. The sites in the third category, identified as infringers, engage in a total violation of intellectual property rights, with the reasons for and consequences of such infringement falling under the purview and discretion of the site's owner and administrator.

In Figure 1, the most frequently consumed categories are detailed according to the indices previously mentioned. These, are divided by the type of material and how users accessed the content: books/E-books/digital magazines, software, film/series/TV, and music.

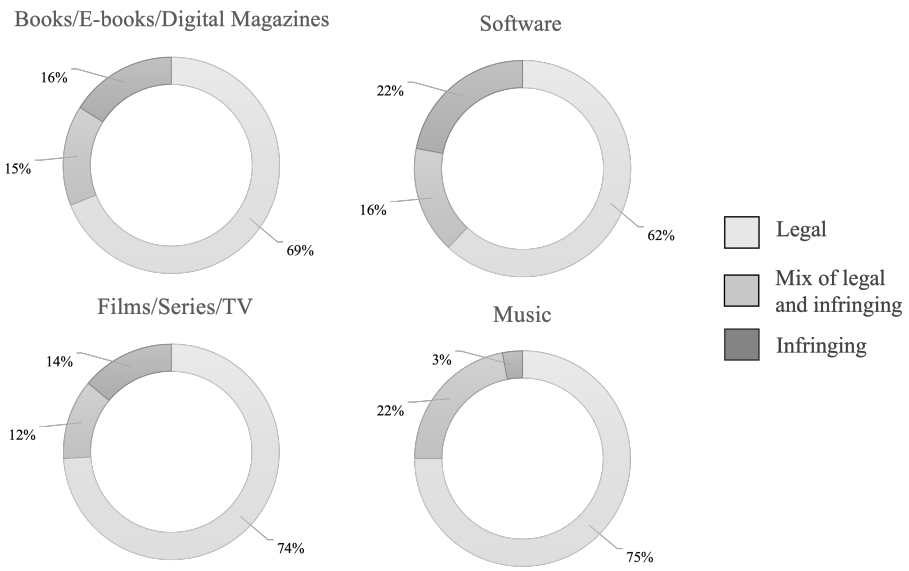


Figure 1. Different categories of multimedia content consumed based on the browsing behavior of a randomly selected group of users, according to the UK Government's Intellectual Property Office [16].

Following this train of thought, there emerges a discernible concern and an overriding need to identify and eliminate materials that infringe copyright in the IoE, through a more efficient and automated process [12,13,15,16]. According to [17,18], in the absence of such measures, the ramifications of these infringements could surpass the mere financial losses currently faced by the industry. These implications include the ethical degradation of work, the introduction of biases in professional training and cultural sectors, a decline in the quality of new creations, and the potential for highly sought-after works to be used as bait in cyber attacks, encompassing malware, phishing, scams, and electronic fraud.

Currently, two primary methods address websites with infringing content over the IoE: legal removal requests and technical tools for scanning and flagging content [19,20]. Responsibility for this falls to entities like the DMCA and the World Intellectual Property Organization (WIPO) [21], who assess these requests. Despite this, the high volume of

complaints has led to the rise of private monitoring services, whose detection techniques are often undisclosed due to privacy concerns [22].

Unfortunately, the policies of Internet Service Providers (ISP) can complicate the removal of infringing content, necessitating monitoring for early detection [23]. Collaborative efforts across industries and academia are developing solutions like browser extensions to deter visits to infringing sites, perimeter security rules to block malicious links, whitelists, and digital watermarking, focusing primarily on prevention scopes [24–26].

Although automated searches in this field have not yet fully matured, the value of Artificial Intelligence (AI), Machine Learning (ML), and particularly Deep Learning (DL) as promising alternatives is increasingly being recognized. The predictive capacity of these technologies to identify patterns indicative of copyright infringement in multimedia content is prematurely being explored [27]. Indeed, current research focuses on unauthorized video sites, academic plagiarism, and unlicensed software, marking a new frontier in proactive exploration for the detection of sites already engaged in infringement.

Even so, preventive, active, or automated strategies are not yet standardized or uniformly structured, thus creating a significant gap for authors who seek to protect their digital multimedia works or to ascertain if these have been disseminated without their authorization. From this consideration, two research inquiries of crucial importance are raised:

First Research Inquiry: In the context of the diverse and varied existing methodologies for safeguarding and discerning rights-infringing content within the realm of the IoE, how could a meticulously structured taxonomy aid authors in making informed and pertinent decisions within their specific contextual framework?

Second Research Inquiry: Within the evolving sphere of IT and acknowledging the scarcity of preceding research in the domain of ML, is it feasible to develop an advanced methodology in ML, employing tracking techniques in the IoE and sophisticated search algorithms, capable of analyzing, processing, and categorizing sites containing potentially infringing content?

The present study addresses the previously posed research questions, focusing on two fundamental axes: firstly, the design of a detailed taxonomy in the domains of SafeGuarding and Active, aimed at managing infringing sites in the IoE, and secondly, the proposal of an innovative methodology for identifying websites that host multimedia content with potential copyright infringement. This research delves into a field that, so far, has received limited attention in the scientific literature [28–35]. The main contributions of this work are as follows:

Point 1: This research introduces a novel taxonomy focused on the dual objectives of protecting multimedia works in the IoE and detecting potentially infringing content, developed through a thorough examination of a wide range of scholarly papers and white papers across various digital libraries.

Point 2: This research introduces an advanced automated search methodology, merging web navigation analysis with multi-engine search capabilities, focusing on key content areas like movies and series, music, software, and books [16]. By harnessing HTML data from websites, the system detects potential copyright violations, ranging from redirects to JavaScript anomalies. It employs BERT (Bidirectional Encoder Representations from Transformers) [36], a state-of-the-art pre-trained encoder, for an innovative synthesis of textual and numerical data. Subsequently, the data are processed through a fine-tuned Dense Neural Network (DNN), marking a significant advancement in information retrieval in the IoE domain. Hereafter, this methodology will be referred to as BERT + DNN.

The remainder of this document is structured as follows: Section 2 outlines the construction of a taxonomy of domains, subdomains, and categories that surround the subject of study—multimedia content and copyright. In Section 3, related work concerning the detection of copyright-infringing material using AI/ML is discussed, delving into their processes and performance. Section 4 introduces the methodological BERT + DNN framework, detailing the stages of data collection, preprocessing, transformation, fine-tuning,

and classification of websites with infringing multimedia content over the IoE. Section 5 presents, compares, and discusses the results using quantitative performance metrics. Finally, Section 6 concludes the paper by listing its significant contributions and emphasizing areas that warrant further exploration in future research endeavors.

2. Taxonomy for the Protection and Detection of Infringing Multimedia Content

In the context of research on copyright infringements within the IoE, multiple studies were discovered, marked by a wide variety of methods, techniques, procedures, and approaches that still lack standardization. This represents a considerable challenge in defining a starting point for projects that are to be implemented. The developed taxonomy offers significant advantages, such as the ability to identify how the dynamics of copyright protection or infringement manifest, the structuring of the techniques used, maintaining coherence between legal and technical tasks, integrating new paradigms in AI/ML, as well as its role as a tool adaptable to emerging information technologies.

To synthesize the findings, the taxonomic architecture is centered around two primary pillars: the studies that focus on safeguarding digital multimedia materials and those that actively operate by detecting where infringing activity is taking place. The study was directed using a highly effective process for conducting searches of documents about the state-of-the-art, known as Systematic Reviews and Meta-Analyses (PRISMA) [37]. This method allows for the establishment of a set of explicit criteria to find, filter, and integrate studies, avoiding redundancy or getting sidetracked by searches that lead to irrelevant outcomes. In Figure 2, the flow of filters used to consolidate the number of relevant records in this research is presented.

To compile an appropriate collection of relevant records, several major scientific publication houses and digital libraries were inspected: EBSCO's Academic Search [38], Taylor and Francis Online [39], Springer Link [40], Elsevier Science Direct [41], Oxford Academic [42], Wiley Online Library [43], Scopus [44], IEEE Explore [45], ACM Digital Library [46], and MDPI [47]. This task was undertaken using search terms such as *digital piracy detection*, *Machine Learning content protection*, *digital rights management with AI*, *copyright infringement*, *multimedia content tracking*, *intellectual property protection*, and *digital content protection*, to name a few.

In summary, a compilation of 44 studies was achieved, divided into 23 within the safeguarding domain and 21 in the active domain, which are detailed in the following paragraphs:

In the realm of safeguarding digital assets, twenty-three significant research works [6,7,19,23,25–27,48–63] have been identified. Of these, nine [6,7,19,23,25–27,48–63] focus on the use of legal resources and partially address technical aspects. Ten of these works [25,26,53–60] adopt a fully technical stance on defensive strategies. However, only four studies [25,26,53–60] are dedicated to protection models that integrate AI and ML-based technologies.

On the other hand, in the active identification of infringing sites, twenty-one [25,28–35,64–75] significant research contributions were compiled. Among these, three [65–67] engage in non-technical actions such as manual search and reporting, contributing to the legal takedown of offending sites. Six projects [28,29,68–71] follow a technical route, involving surveillance and cyber patrolling within the IoE. Regrettably, only twelve initiatives [28–35,72–75] implement one or more AI/ML techniques for identifying infringing content on the IoE, and of these, only eight [28–35] are considered closely related studies. Notably, a single article [25] stands out for its purely technical approach to the defensive and active aspects.

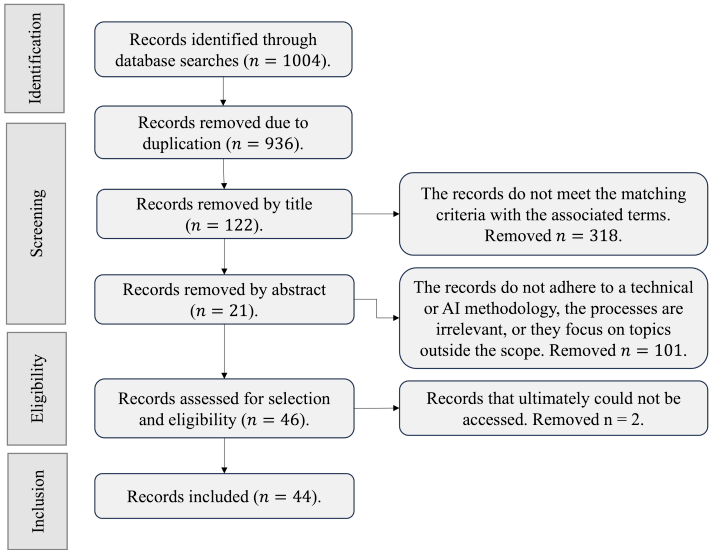


Figure 2. Systematic search filters for records according to PRISMA [37].

Taking into account the defined objectives, proposed hypotheses, abstracts reviewed, methodologies applied, and results achieved, the research is categorized into the following domains and subdomains.

- The *safeguarding domain* [6,7,19,23,48,49] outlines protection strategies related to copyright laws, international and domestic regulations, legal bodies, global and national benchmarks, as well as technical methods implemented to guarantee that multimedia content on the IoE is used only with the explicit permission of the copyright owners. The research identifies the following specific subdomains
  - *Non-technical protection* [50–52] involves examining the legal and non-technical domain of jurisdictional frameworks that govern the presentation and regulation of copyrighted material, ensuring it occurs with the consent of the rights holder.
  - *Technical protection* [25,26,53–60] refers to computational techniques known as digital locks or anti-circumvention mechanisms. These are designed to safeguard multimedia works by digitally enforcing copyright laws.
  - AI/ML-driven protection [27,61–63] is an intelligent scheme that utilizes one or more AI/ML models to enhance the protection of multimedia objects in the IoE, in conjunction with non-technical or technical processes.
- The *active domain* [25,64] are processes and methodologies that actively seek, either manually or automatically, multimedia files across the IoE, to identify, assess, and address potential infringements to which the given object is bound. From these, the following subdomains emerge:
  - *Non-technical* [65–67] contains tasks that reside within the DMCA notice-and-take-down process, an act that sets limitations for multimedia content providers on the IoE. If these mandates are not adhered to, it may result in a partial or total removal of the reb resource.
  - *Technical* [28,29,68–71] involves any technical procedure that allows for traversing the IoE to identify websites, links, URLs, P2P platforms, File Transfer Protocol (FTP) endpoints, torrents, and storage clouds where an identified infringing multimedia resource is residing.
  - AI/ML-driven [28–35,72–75] includes advanced schemes that harness the power of AI/ML algorithms to efficiently perform cyber patrolling by traversing the IoE using NLP techniques, Supervised Learning (SL), Non-Supervised Learning (NSL), and DL analyses to analyze, discover, and present potential infringing multimedia content.

Figures 3 and 4 illustrate the taxonomic structure, covering domains, subdomains, and categories, of the research trajectories within the safeguarding and active domains. Hence, Tables 1 and 2 provide a more detailed view of the aforementioned.

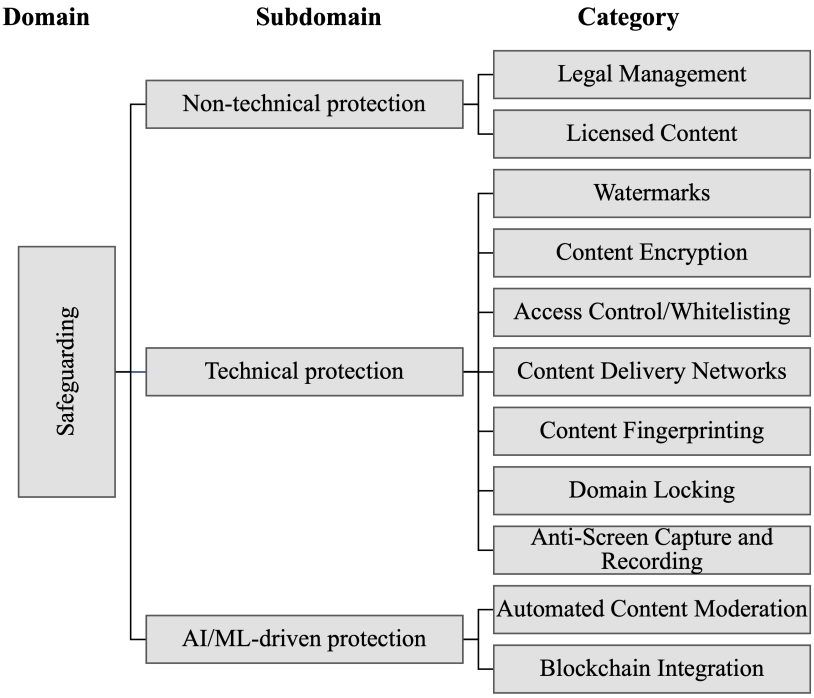


Figure 3. Safeguarding domain, a taxonomy related to safeguarding copyrighted material over the IoE.

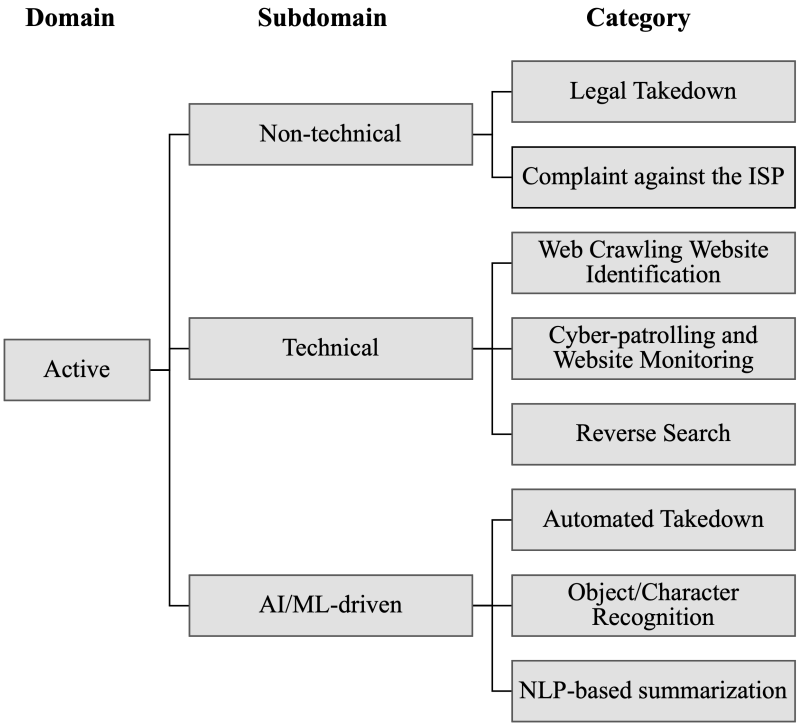


Figure 4. Active domain a taxonomy related to identifying, countering, and removing material that has been infringed over the IoE.

Table 1. Domain,subdomains, and categories related to safeguarding copyright protection of multimedia files in the IoE.

		Category	Description
Safeguarding	Non-technical protection	Legal Management [51]	Methods for legal management of multimedia transmission over the IoE include clauses on authorship, rights, and privacy for official distribution and monetization. It also addresses data collection and analytics practices for the content.
		Licensed Content [52]	Licensing for multimedia on the IoE is defined, confirming the author's legitimacy and the terms of usage, including exclusive licenses, third parties, Creative Commons, EULAs, and public domains. It also covers legal representation and usage warranties.
	Technical Protection	Watermarks [54,61]	In multimedia protection on the IoE, computational procedures like electronic-signatures and anti-tampering steganography ensure against unauthorized changes and piracy. The authenticity of the file can be verified through watermark revelation.
		Content Encryption [55]	Refers to encrypting multimedia files for secure IoE presentation and dissemination. Access is limited to entities with decryption keys, ensuring confidentiality. This also aids in preventing unauthorized access or attacks that compromise their security.
		Access Control/Whitelisting [25,56]	A set of user-centric configurations based on privilege roles and identity, designed to allow, block, or restrict the access and consumption of multimedia content across IoE platforms and websites.
		Content Delivery Networks (CDNs) [57]	These networks efficiently distribute information without directly safeguarding copyrighted content. CDNs' architecture enables synergistic integration of features like watermarks, content encryption, anti-hotlinking, secure transmission protocols (HTTP Strict-Transport-Security and Hypertext Transfer Protocol Secure), constraints on multimedia file distribution, and comprehensive packet monitoring within the network.
		Content Fingerprinting [58]	It is a verification mechanism embedded within the multimedia file. Unlike a watermark, this mechanism exclusively embeds the contractual information of the owner in the form of a hash or digital fingerprint. This can be cross-referenced and verified against intellectual property database records.
		Domain Locking [59]	It is a lock that can be activated by the hosting provider registrant or the CDN to block unauthorized transactions of multimedia content outside the domain. Additionally, it allows for directing the end user to licensed/validated consumption points.
		Anti-screen Capture and Recording [60]	These constitute a series of locks at the operating system level, utilizing control over active directories or through browser-side JavaScript code, to prevent multimedia content from being recorded, trimmed, or copied without the permission of the website owner.
		Automated Content Moderation [62]	Involves the execution of AI, ML, and NLP algorithms to learn about text search behavior and to determine when the terms processed for multimedia content access queries might result in an infringing search. Some of these algorithms are applied using chatbot technology, allowing them to become active blocking tools when an unwanted pattern is detected.
	AI/ML-driven protection	Blockchain Integration [63]	This refers to schemes using AI, ML, and Blockchain to synthesize multimedia files, producing hashes of multimedia databases and catalogs. This facilitates a comparison between the original and infringing files by identifying alterations or removals of metadata or protections, like fingerprinting and watermarking.

Table 2. Domain, subdomains, and categories related to identifying and taking control over multimedia files in the IoE that infringe on copyright.

Active	Non-technical	Legal Takedown [66]	Rights Enforcement Organizations (REOs) address complaints of potential copyright infringements over the IoE. Upon identification by the author or representative, this can lead to legal actions, ranging from takedown notices to lawsuits, compensation, and further prosecution.
		Complaint Against the ISP [67]	Legal tools urging (Internet Service Providers) ISPs and Online Service Providers (OSPs) to adhere to applicable laws within a clear legal framework. Actions can lead to contract terminations with infringers, fines, and penalties for not addressing copyright holder warnings.
	Technical	Web Crawling Website Identification [29,69]	Techniques that navigate through catalogs, explicit requests, databases, or real-time updates of multimedia works using various IoE search engines. These can quantify the similarity of the results yielded from queries with attributes of the retrieved web documents, such as the title, metadata, information embedded in HyperText Markup Language, (HTML), eXtensible Markup Language (XML), JavaScript Object Notation (JSON) dictionaries, hyperlinks, iframes, and URLs.
		Cyber Patrolling and Website Monitoring [28,70]	Using web crawling techniques on the IoE, specific objectives are monitored using robot-type software or web scrapers to track changes on selected sites. Infringement patterns in multimedia files are recorded, and Data Mining procedures evaluate the content to determine the type of infringement.
		Reverse Search [71]	Advanced query configurations on search engines within the IoE aim to match and flag specific multimedia files based on their title, size, metadata, extensions, or other relevant details. These can enhance the identification of infringing websites where the multimedia material is being disseminated. Noteworthy platforms include Google Image Reverse Search [76], TinEye [77], and Yandex [78].
	AI/ML-driven	Automated Takedown [31,34]	It is a novel and experimental technique that employs cyber patrolling using AI, ML, DL, and NLP algorithms to traverse numerous search engines and meta-search engines within the IoE to summarize, classify, and group websites displaying signs of infringement due to possession of multimedia files. It has also expanded to Online Social Networks (OSNs) and the DarkNet [30,73].
		Object/Character Recognition [32,74]	It is a prototype technique that applies Optical Object and Character Recognition (OCR) to determine whether unauthorized advertisements, promotions, offerings, and sales of a multimedia file can be found within the visual content of the website in question. This involves exploring the DOM and converting each image to text for subsequent evaluation.
		NLP-Based Automation [31,33,35,75]	Converts text found within IoE resources into detailed contextual representations, which are designed to detect similarities in potentially infringing content.

3. Related Works

As outlined in Section 2, eight studies [28–35] were identified as closely related works, meeting specific criteria: they employ traversal techniques across the IoE or other related-networks, their methodologies incorporate one or more AI/ML algorithms for identifying websites with potentially infringing multimedia content, and they pertain to key consumption categories as defined in the referenced literature. The following paragraphs provide a detailed overview of each of these publications.

The authors in [28] introduced a methodology to monitor local network traffic, aiming to identify users leveraging P2P protocols for unauthorized content downloads. Utilizing an audio–video fingerprinting system named The CopySense Appliance, that study demonstrated that by inspecting frames within, P2P, TCP, and UDP protocols, traces of non-original multimedia files can be detected, cross-referencing the DMCA database. The findings suggest an imperative for more advanced mechanisms to bolster similarity-based detections.

Moreover, [29] stands as one of the pioneering works leveraging Machine Learning (ML) for infringement content detection. The proposed methodology hinges on metadata

and rule engineering to pinpoint re-uploaded videos on unofficial YouTube accounts. By modifying the Jaro–Winkler distance, this study elucidates the feasibility of measuring the similarity between a query and video metadata, facilitating its classification as either original or infringing.

In contrast, Ref. [30] unveils a technique targeting the unauthorized sale, distribution, and display of various software types. The proposed Ant-Miner trend classification, grounded on the Ant Colony Optimization (ACO) algorithm, is proficient in categorizing the nature of the infringing software.

Expanding on ML applications, Ref. [31] harnesses the prowess of BERT for text summarization. The paper delineates a framework for classifying YouTube videos by modeling topics embedded in brief video descriptions. Such an approach refines DMCA reporting accuracy.

Simultaneously, Ref. [32] emphasizes DL potential in detecting infringements. By deploying DL architectures like AlexNet, ResNet, and a tailored eight-layer DNN, the study successfully identifies re-transmitted banners and logos from streaming services across various platforms.

Taking a web-centric perspective, Ref. [33] champions a web crawling approach coupled with web scraping sensors. The intent is to capture iconography indicative of unauthorized content display or download. Through a marriage of Support Vector Machine (SVM) and Word Embedding techniques, the methodology excels in identifying sites with potentially unauthorized download tendencies based on headers or logos.

In a related vein, Ref. [34] proposes web monitoring through clusters of web scraping sensors. By extracting text, image, and metadata features, a Multi-Tasking Ensemble Algorithm (MTEA) is trained to ascertain unauthorized video streams.

Lastly, Ref. [35] further delves into the capabilities of MTEA for pinpointing unauthorized video streams across diverse platforms. The approach considers chat message ratios and sentiment polarities to gauge a video's popularity vis à vis the original. Complemented by metadata extraction and object recognition, the initial video banner is then compared against the original catalog, earmarking the infringed content.

In order to understand and summarize the capabilities and perspectives of the studies presented in this section, a series of criteria were used, as detailed below:

- Data Collection Mechanism: this aspect evaluates whether the study implements a mechanism for gathering data to acquire samples over the IoE.
- Use of Known Categories: this study is based on a list, white paper, or report that has assessed the approach as being targeted toward a prevalent issue of infringing multimedia content.
- Employment of NLP Methods: it relies in some manner on textual comparison, whether through titles, content, or optical character recognition, to achieve its objective.
- AI/ML Algorithm Utilization: One or more AI/ML algorithms are employed for predictive, regression, or classification tasks.
- Performance Reporting: at least some measure of performance is documented.

Therefore, in accordance with the criteria previously outlined, Table 3 sets out the comparison with the methodology (BERT + DNN) presented herein.

Table 3. State-of-the-art works to compare with the current proposal.

Work	Data Collection over the IoE	Categories	NLP Methods	AI/ML Algorithm	Performance
Copyright Violation on the Internet: Extent and Approaches to Detection and Deterrence [28].	Not reported	Songs and videos accessed without a license over P2P networks.	Vector Space Model-based matching with reported song and video titles by the Detected Attempt to Transfer Copyrighted Media.	Rule-Based Classification (RBC)	Recall = 91.00%
Playing with machines: Using Machine Learning to understand automated copyright enforcement at scale [31].	Partial	Copyright infringing videos on YouTube	BERT sentence-level transformation for titles and descriptions of videos	Logistic regression (LR)	From 54% to 93% probability that a video will be taken down under DMCA.
Software Piracy Detection Model Using Ant Colony Optimization Algorithm [30].	Not reported	Statutes of the Copyright Act for the Legality of Software	Not reported	Ant Colony Optimization (ACO) algorithm for predicting user preferences in unauthorized software usage	Accuracy = 64.94%
Artificial intelligence for detecting media piracy [32].	Not reported	At the authors' discretion: miscellaneous banners and logos from various web streaming platforms.	Not reported	CNN + custom architecture	Accuracy = 98.43%
Machine Learning-Based Advertisement Banner Identification Technique for Effective Piracy Website Detection Process [33].	Yes	Sites suspected of infringing content or piracy based on Alexa Rank and the Google Transparency Report	OCR + Word2Vec	SVM	Precision = 95.00%; Recall = 95.00% and F_1-score = 95.00%.
Crowdsourcing-based Copyright Infringement Detection in Live Video Streams [35].	Partial	Copyright-infringing livestreams on YouTube.	Sentiment polarity in chat message sequences on livestream videos on Youtube via Maximum Likelihood Estimation	AdaBoost, XGBoost, Random Forest (RF), Linear Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).	AdaBoost with 91.03% Accuracy
Copyright infringement detection of music videos on YouTube by mining video and uploader meta-data [29].	Partial	Copyright-infringing channels on YouTube.	Word-level comparison in video title.	Custom classifier based on linear features.	94.68% accuracy in copyright-protected videos.

4. Materials and Methods

In the current study, the methodology used is grounded on the principles of the Cross-Industry Standard Process for Data Mining (CRISP-DM) [79]. CRISP-DM aids in understanding the project's direction, exploring and navigating the data, identifying key factors for project success, and avoiding the repetition of unnecessary phases. Figure 5 illustrates the steps for its development.

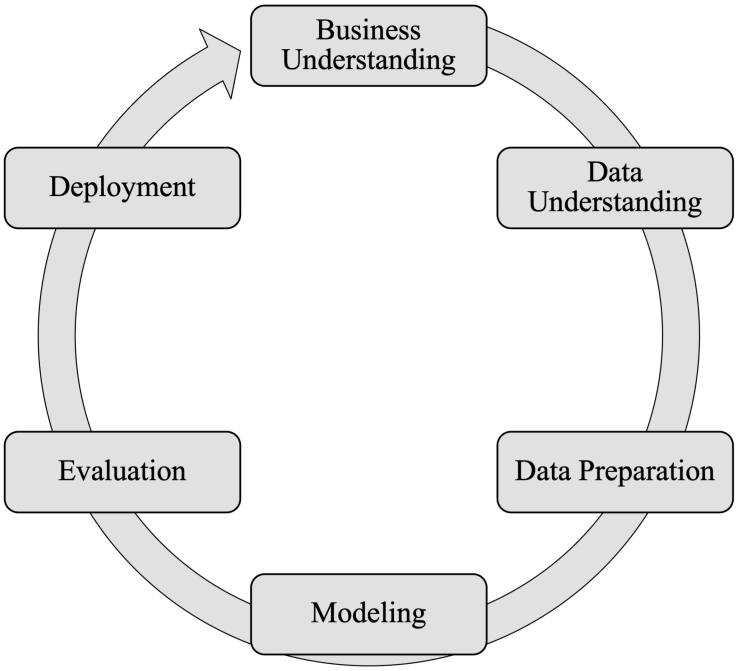


Figure 5. Proposed work methodology based on the CRISP-DM standard to identify sites with infringing multimedia content [79].

In the subsequent Sections 4.1–4.4, each step informed by the CRISP-DM framework is contextualized to fulfill the objectives of this project. It should be noted that this project is methodological in nature; therefore, the Deployment phase is omitted, which is referenced in the list of future work presented in Section 6—Conclusions.

4.1. Business Understanding

In alignment with the CRISP-DM methodology, the process commences with a critical understanding of business needs, leading to a comprehensive series of actions aimed at grasping the project's objectives. Initially, an exploration of the advantages and disadvantages associated with the active domain category—as delineated in Section 2—is undertaken, the specifics of which are encapsulated in Table 4.

Subsequently, the focus shifts to discerning the significance of the identified shortcomings within the non-technical and technical subdomains, thereby revealing the pivotal insights and potential that AI/ML-driven strategies possess for orchestrating a project of this magnitude.

In the ensuing discourse, the investigative intent is accentuated: to rigorously pursue infringing multimedia content by leveraging web crawling techniques across a multitude of search engines within the IoE. The resulting dataset will be engaging advanced NLP algorithms such as BERT and DNN, expanding over a proposed catalog encompassing movies and series, music, software, and books.

Table 4. Advantages and disadvantages of the subdomains of the active domain for the search for multimedia infringing content over the IoE.

Subdomain	Advantages	Disadvantages
Non-technical	• Digital multimedia works that have been flagged for illicit dissemination are decisively taken down, resulting in both personal and monetary compensation to the author. • An anti-piracy and infringement culture is enforced due to the legal actions that may arise if not adhered to. • Laws, standards, and regulations protecting authors are strengthened. • Direct consequences are imposed on circles related to infringement, from ISPs and OSPs to the entity that disseminated the infringing content. • A culture of support and respect for works protected by copyright is fostered, potentially leading to a safer ecosystem within the IoE.	• Infringement notices can be time-consuming, legal trials can be prolonged, and bureaucracy is often extensive, with not all entities being eligible. • Search errors can occur, jeopardizing the legal case. • The discovery of many works is limited in scalability, necessitating manual searches for each, following the same legal pathway.
Technical	• Web crawler robots and scrapers can be customized and can deploy high-range automated search operations to traverse catalogs and databases, where lists of multimedia works intended for exploration on the IoE are presented. • Cyber-patrol systems monitor online illicit multimedia activities in real time and collaborate with legal entities, facilitating automation and prompt takedown actions.	• Web crawlers and cyber patrolling scan the IoE using search engines. Comparing catalogs can lead to errors if algorithms are unsuitable, increasing the false positive rate. • Web crawling and cyber patrolling emphasize catalog breadth over deep analysis, potentially leading to extensive resource use and results that require manual verification. • Companies offering monitoring services often do not disclose the algorithms and techniques they use, leading to a costly monopoly. This can result in contract termination when searches prove ineffective.
AI/ML-driven	• Scalability and precision: The utilization of AI/ML techniques offers a higher level of detection accuracy by revealing inherent patterns within websites. This surpasses traditional cyber patrol methods which typically rely on similarity or distance measures. • Continuous learning and adaptability: In evolving scenarios, AI/ML fosters continuous learning from vast datasets, which can be subsequently integrated into the resulting models. Additionally, these techniques demonstrate high adaptability to various data sources, including HTML, JSON, dictionaries, and other structures that may embed infringing content. • Efficient data processing: AI/ML algorithms facilitate quicker classification, clustering, and prediction of samples pertinent to detecting copyright-infringing multimedia content on the web. Multiple algorithms can be employed for diverse tasks within a single scenario. • Performance metrics: The efficacy of these algorithms can be quantified using performance metrics, which gauge their tolerance levels on test samples. This quantification aids in refining the resulting models or optimizing data preprocessing, guiding it towards a desired learning convergence point.	• Data quality poses a significant challenge. Insufficient data or data with high redundancy, such as embedded text from websites, might degrade the performance of AI/ML models. This can lead to issues with model interpretation or overfitting • Tuning inherent parameters within ML algorithms can be resource-intensive, and even after considerable adjustments, the desired performance might not be achieved. Such challenges can arise due to the biases present when working with unstructured data, typical of websites within the IoE

Based on the list of advantages and disadvantages of the previously mentioned subdomains and categories, the application of AI/ML algorithms offers a more effective and potent solution in this research domain. In light of this, the following success key points can be presented:

- Venturing into the expanse of the IoE, a discerning collection of websites emerges, marked by their propensity to host copyright-infringing multimedia content and their distinct deviation from compliant counterparts. This paves the way for the assembly of a robust dataset, encompassing a diverse array of examples and categories (films

and series, music, software, and books) that amalgamate textual features with the intrinsic dynamics of each site.
- The depth-first (DF) [80] pre-processing algorithm enhances traversal through the DOM, effectively capturing text-holding HTML nodes.
- Data harvested through this DF approach are standardized and converted into contextually rich semantic vectors via pre-trained BERT encoding.
- In the final dataset generation phase, BERT encoding is augmented when merged with additional attributes associated with the unique interactions of each website, thereby amplifying the contextual substance of the samples.
- Final samples are subjected to training and evaluation using a fine-tuned DNN, which, unlike other architectures with specific applications, boasts the flexibility to adapt to multiple objectives, such as classification in this instance.

4.2. Data Understanding

To secure pertinent samples for the training and evaluation of the proposed model (BERT + DNN), key websites listed in the esteemed BrightEdge Top 10 ranking [81] were identified. This ranking is celebrated for its meticulous evaluation, which focuses on recognition, popularity, and traffic within the Internet of Everything (IoE). The selected sites are recognized as the primary platforms where one might encounter notable items within the categories specified by [16]. The identified sites include the following:

1. Movies and series (*MS*) were sourced from the 2022 Golden Tomato Awards: Best Movies & TV of 2022 by Rotten Tomatoes [82].
2. Music (*M*) was based on the 50 Best Albums of 2022 by Billboard [83].
3. Software (*S*) was obtained from the The List of Most Popular Windows Apps Downloaded in 2022, as described by Microsoft [84].
4. Books (*B*) were referenced from Time magazine's 100 Must-Read Books of 2022 [85].

In order to achieve a comprehensive search trough the IoE, a sensor was programmed using Selenium [86], a browser-side automation tool that enables data grabbing by simulating human browsing. Leveraging the capabilities of Selenium, it is possible to capture the HTML content from HTTP protocol responses, to assess the significance of the headers, and to develop specialized storage modules that allow for the data to be saved in comma-separated values (CSV) format files.

The sensor initiates its search using a combination of specific keywords related to each category (*MS*, *M*, *S*, and *B*). For instance, a search might look like *download+anti-hero+Taylor+Swift+rar*. Here, *download* and *rar* are the primary keywords, while *anti-hero* and *Taylor Swift* specify the content being sought. In consequence, an array of auxiliary keywords was considered to combine terms, enforcing more accurate queries: *free download*, *full album*, *full book*, *free crack*, *download*, *online free*, *rar*, *zip*, *compressed*, *7z*, *direct link*, *unlocked*, *serial key*, *free links*, *online watch*, and *direct links*.

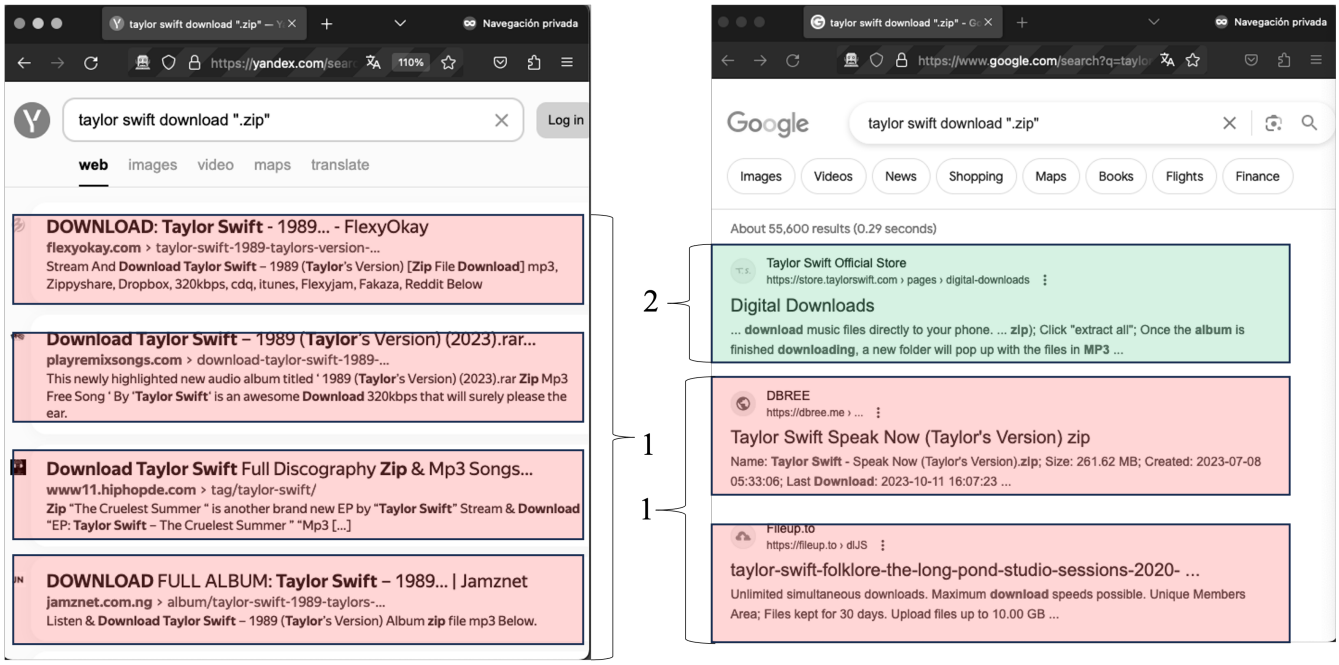
There are several search engines available to navigate a portion of the IoE; in this study, only the five most popular according to The Searching Journal [87] were used: Google Search [14], Yandex [78], Yahoo [88], Bing [89], and DuckDuckGo [90].

The sensor code was outfitted with a generic user-agent header from the automated versions of Firefox and Chromium, with fields that ensure navigation only to the main pages of the previously mentioned search engines, in order to retrieve websites exclusively in American English. In an effort to avoid overloading and mistaking access to already identified non-infringing (NI) websites, a compilation of 17,800 benign URLs was adopted as a whitelist. These URLs are recognized for their impeccable reputation based on the evaluation metrics set forth by Netcraft [91]. This platform assesses the credibility of websites using a diverse set of benchmarks: visitor count; spam list inclusion; desired cybersecurity features; restrained information; and most importantly, whether they have been the subject of takedowns.

Adopting this method, the sensor was constrained to track 50 items for each category, browsing until reaching a cap of 30 result pages on each search engine. It is important to

note that the number of results per page varies by engine: Google fluctuates between 10 and 12 sites, depending on whether they include sponsored links or if the search is related to sales, images, live streams, or news; Yandex consistently displays a steady 10 results per query, regardless of its type; Yahoo provides 5 entries, with the top two links highlighted for their organic relevance; Bing shows between 8 and 10 links sorted by their relevance to current events, social media, and marketing; and DuckDuckGo presents 12, reserving the top 2 spots for the most relevant sites, also featuring links with strong organic positioning.

After obtaining the results, the sensor proceeded to access each of the links, capturing the already-aforementioned features. Within the content context, the title and text embedded in the website's DOM were extracted. Regarding architectural and behavioral attributes, an initial manual evaluation of several potentially infringing sites was conducted, leading to the conclusion that many share key features to be considered: intrusive advertising, questionable reputation, presence of adware, use of URL shorteners, an abundance of scripts run through JavaScript, CAPTCHAs or access puzzles, frequent redirects, numerous cross-domain and download links, as well as iframes. To illustrate, Figure 6 displays some examples of records obtained from the search using a specific combination of keywords, and similarly, Figure 7 details the structure of a website with behavioral patterns that suggest the presence of potential unauthorized multimedia content.



[1] Links that may contain infringing content according to the search criteria during the execution of the Selenium crawler.
[2] Links that were discarded for being considered non-infringing according to the proposed whitelist.

Figure 6. Example of results yielded by an identical query through the web crawling sensor on the search engines Google Search and Yandex.

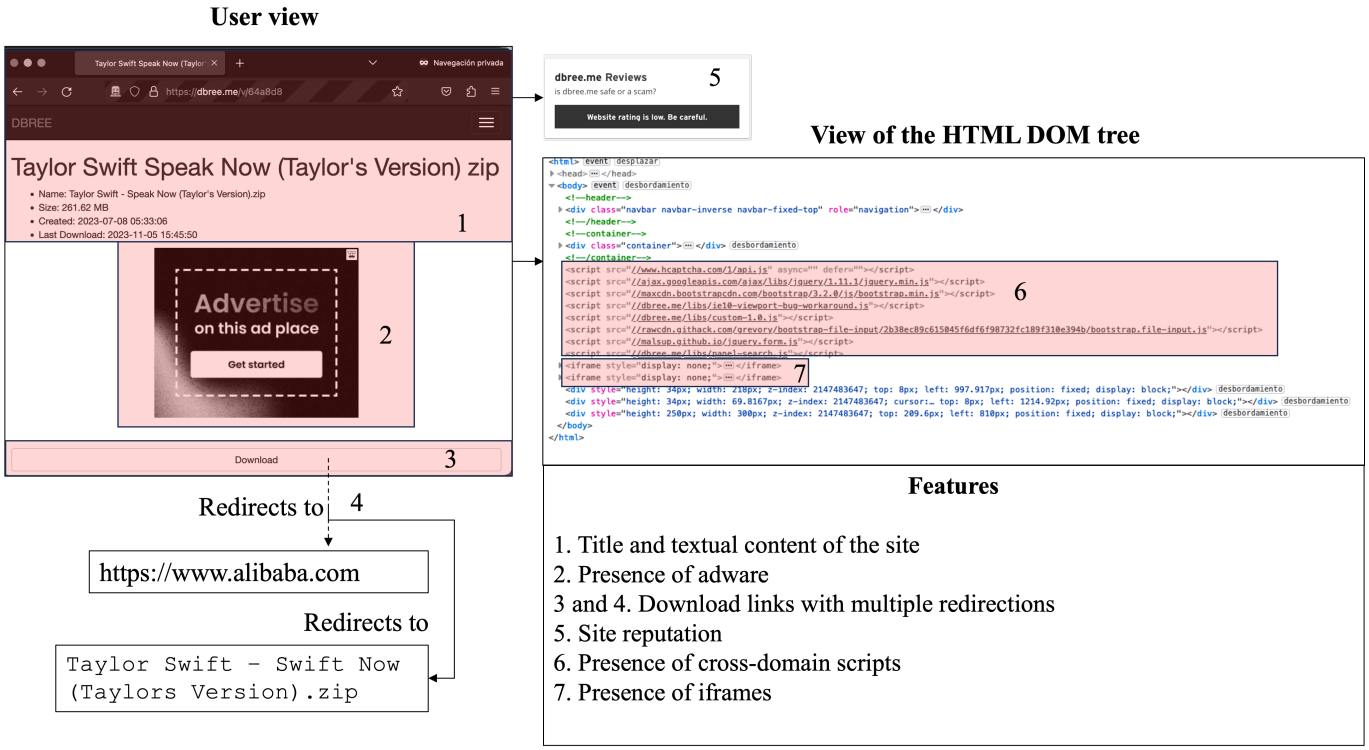


Figure 7. Example of the anatomy of a website with potentially infringing multimedia content.

Upon interaction with the sensor, it became evident that there was a need to add an additional tool known as seeker, integrated into the Mechanical Soup library of the Python Programming Language. This tool aids in tracing the internal links of a website, thus revealing the final destinations hidden by URL shorteners and iframes. The advancement has enabled the discernment of various types of sites that could potentially be in violation. Some, of a less professional nature, promote direct downloads; others distribute content through well-known cloud storage platforms; some utilize their site as a channel to generate income through paid advertising; yet others are plagued with adware, which complicates tracking and analysis. Thereupon, an additional component called Access Level has been added to the toolkit, which classifies how each site presents the pattern of multimedia content distribution that may infringe copyright.

1. *Level* 1: this level included sites that allowed for the downloading of infringing content directly from the homepage.
2. *Level* 2: the download is located on external sites, typically from massive download repositories or personal storage links such as Google Drive, DropBox, and OneDrive, among others.
3. *Level* 3: downloads are offered through external sites, often involving URL shorteners, tracking, or survey sites before the actual link is displayed.
4. *Level* 4: downloads exhibit the characteristics of Level 3, but additionally include challenges, counters, and CAPTCHAs to disclose the displayed content. Many of the sites at this level show signs of adware, poor reputation, or malware components.

In Table 5, the set of features used to construct the dataset is described.

Table 5. Set of features suggested in this project.

Feature	Description	Datatype
Presence of redirects	Presence of forced off-domain redirects. A value of 1 is assigned if present, 0 otherwise.	Binary
Site reputation	Site reputation as determined by Scam Adviser [92], a sensor that gauges the site's quality based on evaluations of pirated content, phishing, or fraud on a scale from 0 to 100. A value of 1 is added if the reputation is below 60; otherwise, 1 is added.	Binary
Adware	A value of 1 is added if any of the links contained on the site are detected by VirusTotal [93] with a detection rate of over 60% as malicious by its sensors. Otherwise, a value of 0 is added	Binary
Presence of URL shorteners	Presence of links with shorteners: a value of 1 is added if one or more links utilize this technology; otherwise, a value of 0 is added.	Binary
Presence of out-of-domain Javascipt content	Presence of scripts coded in JavaScript that are cross-domain. Presence of scripts programmed in JavaScript that operate cross-domain. According to [93], these may be indicative of forced downloads and redirections, the execution of WebAssembly to dissuade the user from remaining on the site, and the initiation of cryptojacking tasks. A value of 1 is added if they are present, and 0 if they are not.	Binary
Presence of CAPTCHAs	In multimedia file download contexts, CAPTCHAs have three main roles: to shield hosts from download abuse, to bolster user confidence in completing the download, and to obscure content tracking by compelling crawlers to solve the puzzle. A value of 1 is added if CAPTCHAs are present, and 0 if not.	Binary
Number of outbound and download links and iframes	Number of download links on site, ifames, redirection links, or shortening links.	Integer
Access level	Proposed additional level based on the patterns observed by the web crawling sensor.	Discrete $[1-4]$
Title	The text title is included as a key feature since it reveals the website's objective.	String
Textual content	The textual content, found within the website's DOM body tags, reveals the site's nature. Content from JavaScript, JSON, and XML tags, or any non-visual elements, was excluded using regular expressions.	String
Label	The labels provided according to the classification perspective are 0 for non-infringing, 1 for movies and series, 2 for music, 3 for software, and 4 for books.	Discrete $[1-4]$

4.3. Data Preparation

The sensor was able to traverse 51,340 websites pertaining to the IoE; however, some conflicts were encountered: 41 sites were blocked due to anti-DDoS protection, 512 returned HTTP 404 (not found) codes, and 607 returned empty content when an attempt was made to download the HTML payload. Consequently, the final dataset encompassed 50,180 samples, of which 17,800 were non-infringing (including the catalog provided by NetCraft), with the

remaining pages potentially infringing: 8780 for the movies and series category, 8401 for the music category, 10,076 for the books category, and 5123 for the software category. In total, 32,380 sites with potential infringing multimedia content were aggregated.

Due to the number of samples and the type of features, there are two main areas to consider in the pre-processing stage. Firstly, the dataset contains binary values (presence of redirects, site reputation, adware, presence of URL shorteners, presence of out-of-domain Javascipt content, and presence of CAPTCHAs), discrete values (level), integer data (number of outbound and download links), and character strings (title and textual content), necessitating transformations and encodings to ensure each feature is on the same scale and context. Secondly, the features must be merged to preserve the necessary latency and to avoid under- or overfitting issues due to bias and variance problems.

To achieve the aforementioned, the features' title and textual content were pre-processed by removing HTML tags within the DOM tree, which presents a challenge since most libraries tend to erroneously eliminate significant portions of textual content. To address this, DF is adopted, which examines the tag graph and identifies nodes with content rich in words, marking them as visited once traversed. The algorithm delves deeper, marking both textual and irrelevant nodes through backtracking until no further elements remain to be explored. The result is a compilation of embedded text. Figure 8 provides a straightforward representation of how the DF approach locates embedded text within the HTML DOM of this project's dataset.

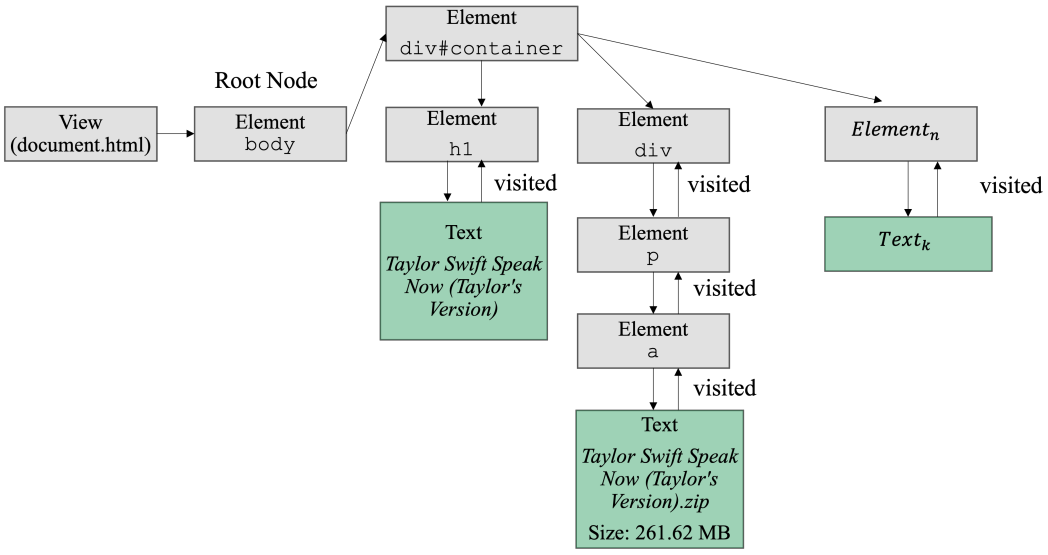


Figure 8. Demonstration of DF traversal in the DOM for the search for embedded textual content.

Ultimately, before adding the title and text as plain text to the dataset, unnecessary white space, emoticons, and any characters outside of the American English UTF-8 encoding were removed, which can then be ingested and transformed by NLP techniques.

4.4. Modeling and Evaluation

This project is marked by the implementation of a fundamental two-phase architecture to compose the BERT + DNN architecture: Initially, contextual vectors are pre-trained, followed by refinement for classification applications. The initial phase is grounded in the use of a pre-existing BERT model, which according to [94], leads the way in advanced understanding of the dynamics between words and their context, based on a diverse textual corpus. As a cornerstone, $\text{BERT}_{\text{uncased}}$ [95] was chosen, notable for its meticulously adapted Masked Language Model (MLM) designed to analyze lowercase English sentences, equipped with a broad and precise vocabulary. This model, structured with 12 layers of attention and 12 attention heads, along with a 768-dimension vector per sequence,

accumulates a total of 110 million parameters, enhancing the adaptability of its parameters for various ML tasks.

The second phase focuses on customizing the embedded representation through the $\text{BERT}_{\text{uncased}}$ model, aligning it with the vocabulary derived from HTML tags. This process encompasses the re-calibration of the model's initial weights and the creation of new vectors, which form the foundation for a fine-tuned DNN, culminating in the development of classification model for the previously mentioned categories: *MS*, *M*, *S*, and *B*. Figure 9 presents a detailed depiction of the proposed architecture.

4.4.1. Pre-Training $\text{BERT}_{\text{uncased}}$

In its pre-training phase, the $\text{BERT}_{\text{uncased}}$ model processes extensive textual data, notably from BookCorpus and Wikipedia, amounting to roughly 250 million words. This corpus undergoes an analysis through three Input Embeddings (E) for Next-Sentence Classification (NSC) and MLM. In NLP, token embeddings methodically segment text into units corresponding to individual words. Positional embeddings assign each token a unique position, integrating this into the embedding layer for improved contextual understanding. Furthermore, segment embeddings are employed to discern between different text sequences, an essential aspect for accurately processing and interpreting linguistic data.

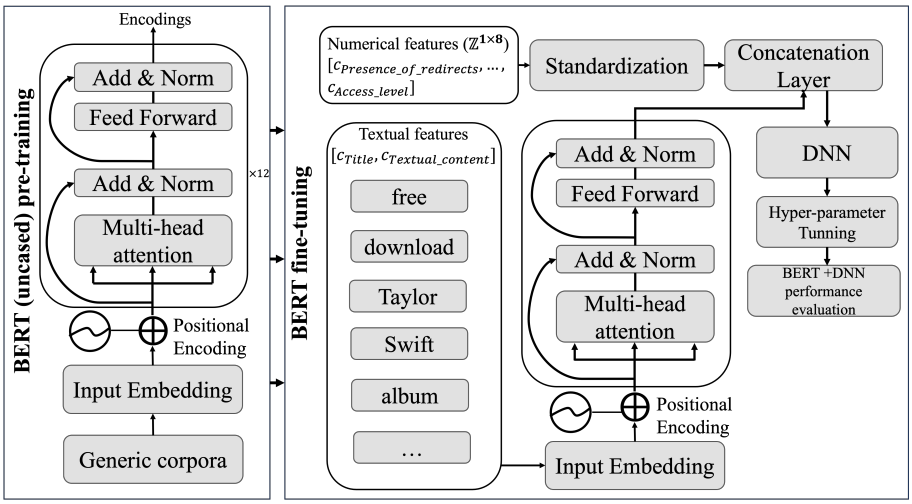


Figure 9. Schematic of architecture for classifying categories *MS*, *M*, *S*, and *B*, utilizing a pre-trained BERT model for subsequent fine-tuning.

To comprehend the mechanism by which $\text{BERT}_{\text{uncased}}$ derives weights from E, an examination of the components constituting the basic BERT architecture is delineated in the subsequent enumeration:

1. *Concatenation with Positional Encoding*: E undergoes integration with the Positional Encoding (PE) layer, thereby infusing information regarding the sequential positioning of each token, resulting in $E^{'} = E + PE$. In this formulation, $E^{'}$ symbolizes the resultant sequences post-integration, embodying a composite of the tokens' semantic information and their respective positions within the sequence.
2. *Multi-Head Attention* (MHA): E' is input into an attention layer for syntactic and semantic analysis of sequences, capturing the language's context and complexity. This is facilitated by the Multi-Head Attention (MHA), using matrix-weight tuples. Query matrices Q identify focus tokens in $E^{'}$, while key matrices K cover all tokens, crucial for computing attention weights and enabling query comparison. Value matrices (V) aggregate outputs from Q and K interactions. Each of the $h_{i=1}^{12}$ heads applies dot-product attention to Q, K, and V, generating scores transformed into probabilities via a softmax function scaled by $\sqrt{d_K}$ (dimension of K). This scaling stabilizes the softmax

function during weight re-calibration in the W^Q, W^K, and W^V matrices, preventing gradient reduction. See Equation (2) for more details.

$$\text{attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^T}{\sqrt{d_K}}\Big)V \tag{1}$$

Consequently, each head $h_{i=1}^{12}$ computes and concatenates the attention results $h_{i=1}^{12} = \text{attention}(Q_i, K_i, V_i)$, as detailed in Equation (2).

$$MHA(Q, K, V) = \text{concatenation}(h_1, \ldots, h_{12})W^O, \tag{2}$$

where W^O is the weight matrix responsible for rescaling the outputs of the concatenation layer towards the subsequent stages of the process.

3. *Add & Norm*: the outputs of the MHA are subjected to a residual connection and normalization prior to entering the Feed Forward (FF) layer. This is integral in any Artificial Neural Network (ANN) to ensure that weights generated during training maintain their significance. This phase emphasizes two crucial processes: addition and normalization.
   Addition is crucial in mitigating gradient vanishing, as it preserves input weight information, denoted as I, across the network. This preservation is maintained irrespective of subsequent layer transformations $F(I)$, achieved by summing the input and transformed output, resulting in the residual output $R_O = I + F(I)$. Subsequently, the role of normalization is to standardize R_O. This standardization is accomplished using the mean-variance normalization method, detailed in Equation (3).

$$R_O^{'} = \frac{R_{O_i} - \mu}{\sqrt{\sigma^2 + E}} \tag{3}$$

   In this process, R_{O_i} denotes the i-th value of the input R_O, with μ and σ^2 representing its mean and variance, respectively. E is the stability constant, and $I^{'}$ is the standardized output. $R_O^{'}$ then undergoes a Linear Transformation (LN), setting the stage for normalization, which optimizes the adaptation of each transformation. See Equation (4).

$$LN(R_O^{'}) = \alpha R_O^{'} + \theta, \tag{4}$$

   where α and θ are the parameters to which the LN transformation is subjected to improve the coupling of $R_O^{'}$ each time it enters the layer.
4. *Feed Forward* (FF): this layer processes values from MHA and *Add & Norm* layers, capturing complex textual characteristics. It employs linear and nonlinear transformations for feature engineering, focusing on key aspects of embedded vectors. These operations occur within $\text{BERT}_{\text{uncased}}$'s dimensions ($\mathbb{R}^{1\times768}$) and its FF network ($\mathbb{R}^{1\times3072}$), concluding with the final *Add & Norm* layer where the model is evaluated, weights are frozen, and pre-training ends.

4.4.2. Fine-Tuning the BERT Model for Classifying Infringing Sites

The refinement procedure is conceptualized as the incorporation of an additional layer to the $\text{BERT}_{uncased}$ transformer model, to construct the BERT + DNN architecture. In this phase, the weights are unfrozen and employed to encode the textual samples from the dataset X. Subsequently, the process involves training, optimizing, and evaluating the DNN estimator for the classification of classes $y = \{MS, M, S, B\}$. Algorithm 1 outlines the steps involved in this task.

For a better understanding of Algorithm 1, the following paragraphs detail the refinement model.

Algorithm 1 Fine-tuning and optimization of the BERT + DNN architecture

Inputs: dataset X, BERT$_{uncased}$ model
Output: BestModel (BERT + DNN)
Load pre-trained BERT$_{uncased}$ model
Unfreeze all 768 weights in BERT$_{uncased}$ model
Pre-process textual features X_{text} via BERT$_{uncased}$(tokenization, padding) model
Preprocess numerical features X_{num} via Standard Scaling
for each data point $(X_{\text{text}_i}, X_{\text{num}_i}) \in X_T$ **do**
    Generate vector $\mathbf{x}_i$ from X_{text_i} using BERT$_{uncased}$ model
    Concatenate $\mathbf{x}_i$ with X_{num_i} to form $\mathbf{x}_{\text{concat}_i}$
end for
Construct the concatenated training subset $X_{\text{concat}} \leftarrow \mathbf{x}_{\text{concat}_i^n}$
Partition dataset X_{concat} into training $X_{T_{\text{concat}}} \in X_{\text{concat}}$ (80%) subset and validation subset $X_{V_{\text{concat}}} \in X_{T_{\text{concat}}}$ (20%)
Initialize the DNN with Input size $I = 776$
Initialize the DNN with output size $O = 5$
Define the subset $\mathcal{P}$ hyperparameters an values to optimize: learning rate η, epochs $\mathcal{E}$, batch size $\mathcal{B}$ and number of hidden layers H_M with their respecipve units U
Initialize BestModel (BERT + DNN), BestPerformance $\leftarrow$ null, 0
for each subsetset of hyperparameters ($\mathcal{P}$) **do**
    Compute the Excepted Improvement function $EI(\mathcal{P})$ on BERT + DNN using $(X_{T_{\text{concat}}}, y_T)$
    Compute Sparse Categorical Cross-Entropy loss function ($\mathcal{L}$)
    Backpropagation to update BERT + DNN weights
    if performance of $EI(\mathcal{P})$ on $(X_{V_{\text{concat}}}, y_V)$ > BestPerformance **then**
        BestModel $\leftarrow$ BERT+DNN
        BestPerformance $\leftarrow$ performance of BERT +DNN
    else Select the next subset of hyperparameters $\mathcal{P}_{\text{next}}$ to improve $EI(\mathcal{P})$
    end if
end for
return BestModel (BERT + DNN) with the best subeset of hyperparameters $\mathcal{P}^+$

The parameters of the BERT$_{\text{uncased}}$ model are initialized and subsequently unfrozen, following the recommendations of [96], for large-volume databases. This approach is suggested as it allows for the re-calibration of these parameters as they adapt to the dataset X, with the aim of achieving a more efficient adaptation to the classification task.

The textual features *Title* and *Textual Content*, incorporated into X as X_{text}, are processed through a refined procedure of tokenization and subsequent alignment to standardize the length of the tokens before their masking. For this purpose, a maximum length of 300 fixed elements per token has been selected, applying a uniform padding function to complete those sequences that do not reach this maximum limit. The latter generates an embedded output vector code of dimensions $\mathbb{R}^{300\times768}$ for each textual sample.

The numerical samples $X_{\text{num}} \in X$ which include variables such as *Presence of URL shorteners*, *Presence of out-of-domain Javascript content*, *Presence of CAPTCHAS*, *Number of outbound download links and iframes*, and *Access level*, undergo to a Standard Scaling (SC) operation. This process involves removing the mean and adjusting each of them to a unit variance, a process detailed in Equation (5).

$$X_{\text{num}_i} = \frac{X_{\text{num}_i} - \mu_i}{\sigma_i}, \tag{5}$$

where X_{num_i} represents each of the numerical samples, μ_i is the mean, and σ_i is the variance.

The textual embedded vectors $\mathbf{x}_i$ from X_{text_i} corresponding to the i-th position are merged with their respective standarized numerical vectors X_{num_i}, culminating in the

creation of a final set X_{concat} that harmoniously integrates both characteristics, as detailed in Equation (6).

$$X_{\text{concat}} = \sum_{i=1}^{n} (\mathbf{x}_i \oplus X_{\text{num}_i}) \tag{6}$$

X_{concat} is then conformed by $\mathbf{x}_i \in \mathbb{R}^{i \times 776}$ samples, composed of fixed-size vectors of 768 elements from the embedding obtained from BERT$_{uncased}$ and the eight remaining standardized numerical features. Afterwards, X_{concat} is divided into three different subsets: $X_{T_{\text{concat}}}$, which is used to train the BERT + DNN algorithm; $X_{P_{\text{concat}}}$, intended for performance testing; and $X_{V_{\text{concat}}} \in X_{T_{\text{concat}}}$, which is used for accuracy verification during training. This last step is essential for adjusting the weights of BERT and the hyperparameters of the DNN. These subsets represent, respectively, 70%, 30%, and 20% of the total and are selected randomly and without replacement.

At this stage of development, BERT + DNN has been configured to receive samples and to initiate the training process. This phase involves the integration of DNNs [97], classified as a category within DL algorithms, distinguished by their proficiency in a range of activities including synthesis, classification, and outlining of concepts in unstructured data, such as those found on websites. The DNNs have solidified their robustness, particularly in operations related to the analysis of information from the IoE, as a case in point, the detection of Denial of Service (DoS) attacks, the identification of internal/external cybersecurity threats, the discernment of web attacks through false data injection, the neutralization of cyber assaults, the containment of phishing attempts, the discovery of piracy activities, and the prevention of the proliferation of malicious software, to name a few.

Various architectures of DNNs exist, distinguished by the quantity of dense hidden layers, the types of activation functions utilized, the optimizers employed to reduce error, and the number of units present in the output layer. Despite the plethora of configurations, the architecture delineated in [98] has shown exceptional efficacy in the detection of cryptojacking patterns within website payloads when juxtaposed with other neural network models like Convolutional Neural Networks (CNNs), RNNs, and LSTMs.

The BERT + DNN model starts with 776 units for processing the input and 5 for the output, each output representing a different class ($y = MS, M, S, B$). In this method, adjustments are made to several hyperparameters, including the number of epochs ($\mathcal{E}$), the learning rate (η), the batch size ($\mathcal{B}$), and the number of layers (H), along with the number of units in each, forming these adjustments into a series of subsets that integrate into the set $\mathcal{P}$. The process is iterative and seeks to optimize these parameters through a function that maximizes the Expected Improvement (EI), with the goal of achieving the best possible accuracy, namely the $BestPerformance$. The identification of this optimum, $BestPerformance$, is carried out by evaluating X_{concat_V} along with y_V using the best subset of $\mathcal{P}^+$ at the conclusion of the training process, resulting in the selection of the most effective model ($BestModel$).

To find the best set of hyperparameters for BERT + DNN in Table 6, the parameters and ranges used are summarized. It is crucial to highlight that the ReLu activation function is used both in the input layer, I, and in the hidden layers, H_M, while the Sigmoid function is employed in the output layer, O.

Table 6. Summary of the proposed parameters for BERT + DNN, detailing the ranges used for its validation.

Parameter	Range
$\mathcal{E}$	$\{75, 100, 150\}$
$\mathcal{B}$	$\{16, 32, 64, 128\}$
η	$\{1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$
I (Input Layer)	776
H_M (Hidden Layer)	$M = \{1, 2, 3, 4, 5\}$; $U = \{32, 64, 128, 256, 512\}$
O (Output)	5

For stable operation, a BERT+DNN requires the establishment of several foundational configurations, detailed as follows:

- *Batch normalization layer*: this component ensures the stabilization of the dense layer's (H_M) output by normalizing the values, aiming for a standard deviation near one and a mean approaching zero.
- *ADAM (Adaptive Moment Estimation)*-type optimizer: this mechanism efficiently manages the adjustment of each unit's weights, utilizing a learning coefficient in conjunction with stochastic gradient descent. It anticipates the estimation of first- and second-order moments, maintaining low computational complexity throughout the process.
- *Categorical Cross Entropy loss function ($\mathcal{L}$)*: for multi-class classification, where the output layer O employs a Sigmoid function, the loss can be determined by assessing the difference between the target classes y_V and the predicted classes $\hat{y}_V$ from the validation set $X_{V_{\text{concat}}}$, during the training step. The goal is to minimize this discrepancy by using the divergence between the probability distributions of each class $C = 5$, as shown in Equation (7).

$$\mathcal{L}(y_V, \hat{y}_V) = -\sum_{i=1}^{C} y_{V_i} \log(\hat{y}_{V_i}) \tag{7}$$

The formulation of the optimization problem is presented as follows: Let $\mathcal{P} = \{\mathcal{E}, \mathcal{B}, \eta, H_M\}$ be the set of hyperparameters and $f(\mathcal{P})$ be the function aimed at maximizing accuracy. $f(\mathcal{P})$ is modeled through a Gaussian process that calculates the mean of ($\mu(\mathcal{P})$) and its covariance function ($k(\mathcal{P}, \mathcal{P}^{'})$), where $f(\mathcal{P}) \leftarrow [\mu(\mathcal{P}), k(\mathcal{P}, \mathcal{P}^{'})]$. Subsequently, $f(\mathcal{P})$ is subjected to EI, as detailed in Equation (8).

$$EI(\mathcal{P}) = \mathbb{E}[max(f(\mathcal{P}) - f(\mathcal{P}^{+}))], \tag{8}$$

where $\mathbb{E}$ represents the expected value of the outcome of the set exhibiting the highest accuracy, calculated from the difference between the current set $f(\mathcal{P})$ and the one that achieved the best value $f(\mathcal{P}^{+})$.

The training process will continue until all hyperparameters $\mathcal{P}$next have been optimized to maximize the $EI(\mathcal{P})$ function, with validation conducted using the tuple XV_{concat}, y_V. Consequently, the $BestModel(BERT + DNN)$ will be acquired and subsequently employed to evaluate the final performance with $\hat{y}_{P_{concat}} = BestModel(BERT + DNN) \leftarrow (X_{P_{\text{concat}}}, y_{P_{concat}})$.

In the context of multi-label classification, Figure 10 displays a confusion matrix that evaluates the predictive performance of the resultant model $BestModel(BERT + DNN)$ across the selected classes: non-infringing (NI), movies and series (MS), music (M), software (S), and books (B). The matrix reveals two significant values: TP_k (true positives), indicating the number of correct predictions for the C-th class compared to the ground truth, as well as FP_k (false positives).

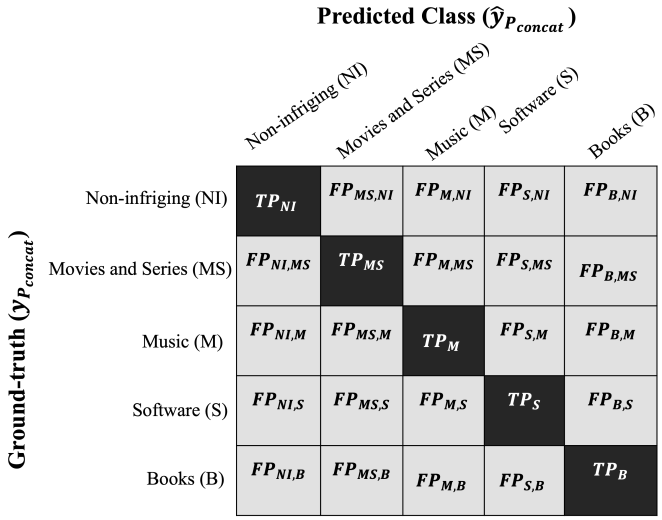


Figure 10. Multi-class confusion matrix for the labels non-infringing (NI), movies and series (MS), music (M), software (S), and books (B).

The multi-class performance metrics used to evaluate the BERT+DNN model are described in Table 7.

Table 7. Multi-class performance metrics used to evaluate the DNN's efficacy, where $i = 1$ denotes the index ranging from the first to the k-th class.

Metric	Description	Mathematical Definition
Precision	It is the fraction of relevant predicted observations for each class with the total number of classified instances with respect to the others.	$Precision_{i=1}^{k} = \frac{TP_{i=1}}{TP_{i=1}+\sum_{i=1}^{k} FP_i}$
Recall	It is the fraction of relevant samples (true positives) predicted for each class concerning the total number of relevant instances.	$Recall_{i=1}^{k} = \frac{TP_{i=1}}{\sum_{i=1}^{k} TP_i}$
F_1-score	The F_1-score represents the harmonic mean of precision and recall, offering a balanced measure of a model's performance by assessing its equilibrium between $TP_{i=1}^{k}$ and $FP_{i=1}^{k}$.	$F1_{i=1}^{k} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

5. Results and Discussion

After applying hyperparameter adjustments to the $X_{V_{\text{concat}}}$ set, validation analyses $EI(\mathcal{P})$ revealed that the most optimal parameters, $\mathcal{P}^{+}$, include an η of 1×10^{-4}; a batch size ($\mathcal{B}$) of 32; a three-layer configuration (H_M) with 32, 64, and 128 units (U) respectively; and a total of 100 epochs ($\mathcal{B}$).

Using the previously specified values, the test tuple ($X_{P_{\text{concat}}}$, y_P) was trained using $\mathcal{P}^{+}$ and subsequently validated with the subset ($X_{V_{\text{concat}}}$, y_V). Accuracy rates of 95.14% during the training phase and 98.71% in the testing stage were recorded. The convergence of the learning process is evidenced in Figure 11.

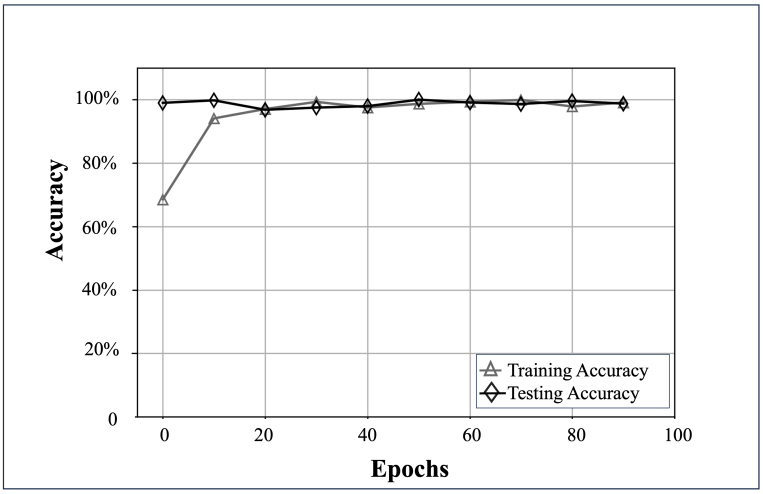


Figure 11. Convergence of accuracy rates throughout the training and testing stages, applying the most refined and optimal set of hyperparameters $\mathcal{P}^+$.

To assess the classification results achieved by the BERT+DNN model, Figure 12 presents the multi-class confusion matrix, along with the support (number of samples submitted) for each test sample.

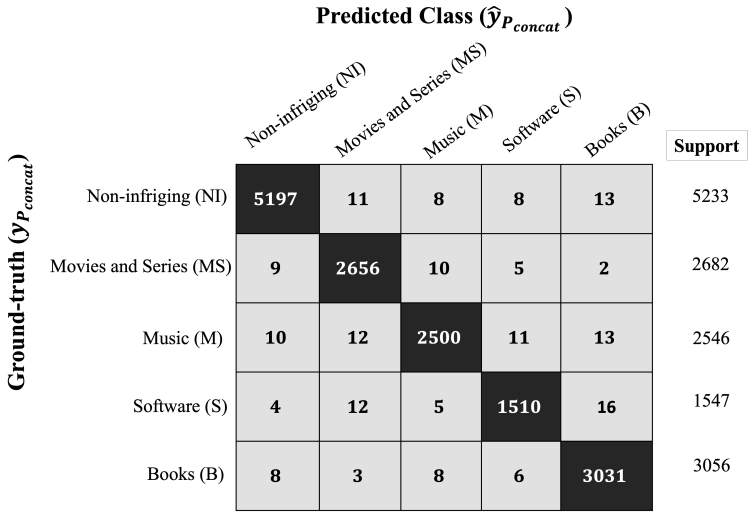


Figure 12. Confusion matrix representing classification outcomes across labels: non-infringing (NI), movies and series (MS), music (M), software (S), and books (B).

Once the composite confusion matrix has been constructed, it is possible to calculate performance metrics, including precision, recall, and the F_1-score, whose values are referenced in Table 8.

Table 8. Performance results of the BERT+DNN proposal obtained for classes NI, MI, M, S, and B.

Class	Precision	Recall	F_1-Score
NI	99.41%	99.31%	99.36%
MS	98.70%	99.03%	98.86%
M	98.78%	98.19%	98.48%
S	98.12%	97.61%	97.86%
B	98.57%	99.18%	98.87%
Macro average	98.71%	98.87%	98.87%

Although it is not possible to directly compare state-of-the-art studies in terms of their methodology, as they do not use a multi-class classification model like the one employed in this BERT + DNN project, we highlight some of the main disadvantages of these works in the following lines. For the sake of clarity and linguistic conciseness, the related works [28–33,35] will from now on be referred to as studies.

Regarding the exploration of the Internet of Everything (IoE), studies [28,30,32] do not conduct a thorough search for multimedia content at risk of infringement. Study [28] narrows its focus to the detection of downloads of multimedia objects that have been previously identified and transmitted via a Local Area Network (LAN) at the application layer, identified by a protocol analyzer that targets Peer-to-Peer (P2P) ports. Meanwhile, study [32] bases its dataset on interviews that investigate the use of potentially infringing software among users of varying ages, genders, ethical beliefs, and economic statuses, without providing details on the aggregation of data over any specific transmission medium. Study [32] utilizes a pre-established database featuring logos highly susceptible to copyright infringement, depicted in images of well-known commercial brands, and conducts experiments based on this dataset, yet it does not describe the process or location within the IoE from which these data were collected.

Contrary to other methods, studies [29,31,35] conduct selective explorations, opting for web scrapers over search engines to analyze specific sites within the Internet of Everything (IoE) that are directly linked to their research objectives. For instance, Ref. [31] directs its scraping towards specific YouTube endpoints to retrieve the top K videos displayed without user consent, solely collecting titles and descriptions. Similarly, Ref. [29] also gathers data from YouTube, but it uses a more advanced and semantically enriched latent vocabulary that categorizes results based on ethically questionable practices for accessing protected content. Meanwhile, study [35] employs a hybrid strategy, interspersing the detection of live streams with compromised copyright with a dataset customized for research purposes.

First comparative finding: Compared to the BERT+DNN methodology, which explores a section of the IoE using five search engines, studies [28,30,32] completely restrict their scope without retrieving any information from the IoE; on the other hand, Refs. [29,31,35] only address a fraction of the IoE and fail to achieve a broader spectrum in their collection due to the absence of search engine utilization. Only Ref. [33] conducts an effective traversal, but its effort is confined to recovering just 98 sites associated with pirated content, a number not considered substantial for a study of this caliber.

In the realm of exploring relevant categories, only the studies [28,33] propose a wider spectrum of content for infringement detection. For instance, Ref. [28] features a collection of music and videos in line with Apple's rankings, while Ref. [33] turns to the now-defunct AlexaNet to identify sites likely to be monitored in their data collection, covering areas such as video games, torrents, and sports. In a more generic approach, Ref. [30] superficially addresses the software category without specifying which programs might be more prone to illicit use. Regarding videos, Ref. [29,31,35] focus on real-time streaming platforms without delving into how to handle TV series, movies, or other streaming content types. In a more limited scope, [4] relies on brands from recognized providers without describing their specific fields of expertise. Studies [28–33,35] fail to incorporate additional features that define the behavior and structure of the sites analyzed.

Second comparative finding: In contrast to previous research, this methodology (BERT+DNN) encompasses a wider range of categories (NI, MS, M, S, and B), delving into a more comprehensive analysis of distinctive features of infringing sites. This analysis includes elements such as redirections, reputation assessment, the presence of malicious elements (adware), the use of URL shorteners, and the improper use of JavaScript, among others, factors that have not been addressed by other studies.

In the domain of NLP methods, studies [30,32] do not utilize textual analysis methods, erroneously assuming that infringing content can be detected based on specific dependencies. This approach is significantly flawed as it overlooks semantic subtleties, potential

issues like case sensitivity, ambiguity, and comprehension limitations, resulting in a final product that cannot be effectively assessed for performance [99].

Conversely, study [28] employs the Vector Space Model (VSM). However, this method also faces challenges, as it has been shown that a pure vector representation lacks semantic depth, contributes to an unstable vocabulary dimension, and leads to data dispersion, yielding vector outputs of low quality [99].

In study [29], queries based on word similarity are utilized, also through VSM. Such comparisons between sentences risk scale sensitivity in the resulting vectors due to their reliance on the frequency and weight of word occurrences, which can lead to inappropriate similarities [100].

Study [33] adopts a more robust representation by combining Optical Character Recognition (OCR) with Word2Vec. Nonetheless, there are risks of mismatch and contextual misinterpretation when using OCR with Word Embeddings, as character misrecognition could lead to incorrectly interpreted words that distort the sense captured by Word2Vec [101]. Additionally, it is established that Word2Vec cannot recognize the order of words across multiple sentences, which limits understanding [102].

Moreover, study [35] proposes extracting sentiment polarity as a feature. This can lead to oversimplification since a user's text may carry multiple meanings depending on the emotional context, potentially misinterpreting the underlying sentiment [103]. Furthermore, these features lack context, as polarity does not provide semantic information, only interpretation.

Only study [31] and the current project employ BERT, which, as discussed throughout the manuscript, reduces the need for extensive feature engineering, preserves the latent context of sentences, and ensures that words maintain their meaning regardless of their position and frequency in the text.

Third comparative finding: In contrast to the previously mentioned research, the methodology presented here (BERT+DNN) harnesses BERT's pre-trained weights to consolidate, abstract, and more accurately reflect the text embedded within websites. BERT has been meticulously engineered to address the myriad shortcomings prevalent in NLP models, issues that have been reiterated in this section: the lack of semantic depth, the intricacies of composition and connections, the challenges posed by high dimensionality and data dispersion [28,29], contextual rigidity [33], and premises inclined to introduce ambiguity [35].

In the cutting-edge realm of Artificial Intelligence and Machine Learning (AI/ML), assessing the effectiveness of various models is critical. Study [28] initially presents RBC, which excels with small and structured datasets. Yet, its rigidity in the face of complex rules diminishes its applicability for the synthesis of textual and numerical features, which is a distinctive feature of this research (BERT+DNN), and restricts its ability to grasp nonlinear dynamics [104].

While study [28] stops short of detailing the RBC algorithm's mathematical underpinnings, the prevailing literature suggests that its linear orientation significantly undermines its utility with nonlinear data types, such as textual content, often leading to error-prone models with subpar performance.

Study [30] acknowledges the intricacy of ACO, a formidable challenge for amalgamating features in multidimensional contexts, necessitating a thorough search methodology to pinpoint websites likely in breach of copyright laws [105].

Moreover, the study [31] endorses LR, an algorithm attuned to predictive tasks involving less complex datasets. However, the process of integrating diverse data characteristics (BERT+DNN) calls for significant regularization trials, potentially detracting from its efficacy in identifying multimedia content within digital lexicons.

Correspondingly, study [32] employs CNNs, renowned for their proficiency in object detection, yet their performance can be erratic within the heterogeneous IoE. A recalibration of the DOM could potentially refine the detection of illicit multimedia material.

Finally, study [35] introduces fundamental algorithms like AdaBoost, XG-Boost, RF, SVM, and MLP. Despite the unique challenges each algorithm faces in pinpointing infringing content, they all share the common requirement for meticulous hyperparameter optimization [106]. AdaBoost risks overfitting without proper classifier baseline selection, XG-Boost may misinterpret features with outlier data, Random Forest necessitates exacting adjustments of its estimators, SVM demands intricate kernel and regularization parameter selection, and MLP grapples with learning rules that could erode the integrity of weight values.

Fourth comparative finding: The BERT+DNN architecture is more versatile and streamlined, effectively managing the fusion of textual and numerical features. It exhibits a classification potential that surpasses the performance metrics of other state-of-the-art studies (precision = 98.71%; recall = 98.67%, and F_1-score = 98.67%) in detecting infringing multimedia content, outperforming most the state-of-the-art works.

6. Conclusions and Future Research Directions

This manuscript has introduced an innovative methodology, named BERT+DNN, which addresses the detection and evaluation of infringing content in the IoE for key categories such as movies and series, music, software, and books. This methodology has greatly benefited from a structured taxonomy, answering the first research question. The taxonomy not only effectively categorizes the content but also provides authors with a crucial tool for making informed and relevant decisions within their specific contextual frameworks. By accurately identifying and classifying different forms of infringing content, authors and stakeholders can adapt their strategies more efficiently and effectively based on the particular challenges they face.

Moreover, in response to the second research question, the feasibility of developing an advanced methodology in ML, using tracking techniques in the IoE and sophisticated search algorithms, has been demonstrated. The BERT+DNN methodology not only analyzes, processes, and categorizes websites containing potentially infringing content but also does so with a precision of 98.71% and a recall and F_1-score of 98.67%. This achievement underscores the viability and effectiveness of applying advanced ML approaches in the fight against copyright infringement in the IoE.

While limitations regarding scope, collection methods, characterization, and the algorithms used have been discussed, these findings open new avenues for future research and development in this field. Thus, the current study not only surpasses existing approaches in the detection of infringing content but also offers a robust framework for informed decision-making and the implementation of copyright protection strategies in an ever-evolving digital environment. Recognizing this as a step in an expansive, under-explored field, this study suggests future research directions:

- Integrating Optical Character Recognition (OCR) for DOM exploration to identify non-explicitly depicted multimedia objects;
- Merging the approach with an API for direct reporting to authorities, enabling automated content takedown;
- Constructing the deployment phase of CRISP-DM by processing URLs for real-time monitoring and classification;
- Broadening the range of categories to encompass piracy, live-streaming, social media networks, and the DarkNet.

Author Contributions: A.H.-S., G.S.-P., L.K.T.-M., H.M.P.-M., J.P.-P., and J.O.-M. contributed equally to this work. 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 available on request from the corresponding author.

Acknowledgments: The authors thank the National Council for Humanities, Sciences, and Technologies (CONHACYT) and the Instituto Politécnico Nacional for their support of this research.

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

References

1. Graham, M.; Dutton, W.H. *Society and the Internet: How Networks of Information and Communication are Changing Our Lives*; Oxford University Press: Oxford, UK, 2019.
2. Roblek, V.; Meško, M.; Bach, M.P.; Thorpe, O.; Šprajc, P. The interaction between internet, sustainable development, and emergence of society 5.0. *Data* **2020**, *5*, 80.
3. Fan, X.; Liu, X.; Hu, W.; Zhong, C.; Lu, J. Advances in the development of power supplies for the internet of everything. *InfoMat* **2019**, *1*, 130–139.
4. Farias da Costa, V.C.; Oliveira, L.; de Souza, J. Internet of everything (IoE) taxonomies: A survey and a novel knowledge-based taxonomy. *Sensors* **2021**, *21*, 568.
5. DIGITAL 2022: GLOBAL OVERVIEW REPORT. Available online: https://datareportal.com/reports/digital-2022-global-overview-report (accessed on 21 October 2023).
6. Liu, J.; Wang, X.; Wang, Y. Research on Internet Copyright Protection Mechanism: Based on the Perspective of the Comparison of Chinese and American Legislation. In Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022), Wuhan, China, 25–27 March 2022; Atlantis Press: Amsterdam, The Netherlands, 2022; pp. 1592–1600.
7. Madi, R.; Al Shamsi, I. A brief overview of the exemptions to the prohibition on circumvention of technological protection measures under the DMCA: Any similar exemptions under the UAE legislation? *Int. Rev. Law Comput. Technol.* **2021**, *35*, 352–364.
8. Kalyvaki, M. Navigating the Metaverse Business and Legal Challenges: Intellectual Property, Privacy, and Jurisdiction. *J. Metaverse* **2023**, *3*, 87–92.
9. Nakamura, S.; Enokido, T.; Takizawa, M. Protocol to efficiently prevent illegal flow of objects in P2P type of publish/subscribe (PS) systems. *Serv. Oriented Comput. Appl.* **2019**, *13*, 323–332.
10. Ku, R.S.R. The creative destruction of copyright: Napster and the new economics of digital technology. *Univ. Chic. Law Rev.* **2002**, *69*, 263–324.
11. Peukert, C.; Claussen, J.; Kretschmer, T. Piracy and box office movie revenues: Evidence from Megaupload. *Int. J. Ind. Organ.* **2017**, *52*, 188–215.
12. Yadav, M.A.; Singh, N.B. The Ineffectiveness of Copyright System to Respond Effectively to Digitalization and Possible Measures. *Spec. Ugdym.* **2022**, *1*, 4531–4537.
13. Google: Copyright Infringing URL Removal from Domains 2022. Available online: https://www.statista.com/statistics/279954/infringing-urls-requested-to-be-removed-from-google-search-by-domain/ (accessed on 21 October 2023).
14. Google. Available online: https://www.google.com/ (accessed on 21 October 2023).
15. McGhee, H. Reinterpreting Repeat Infringement in the Digital Millenium Copyright Act. *Vanderbilt J. Entertain. Technol. Law* **2023**, *25*, 483.
16. Online Copyright Infringement Tracker Survey (12th Wave). Available online: https://www.gov.uk/government/publications/online-copyright-infringement-tracker-survey-12th-wave (accessed on 21 October 2023).
17. Alabduljabbar, A.; Ma, R.; Alshamrani, S.; Jang, R.; Chen, S.; Mohaisen, D. Poster: Measuring and Assessing the Risks of Free Content Websites. In Proceedings of the Network and Distributed System Security Symposium (NDSS'22), San Diego, CA, USA, 26 February–1 March 2022.
18. Bradley, W.A.; Kolev, J. How does digital piracy affect innovation? Evidence from software firms. *Res. Policy* **2023**, *52*, 104701.
19. Foley, J.P. Ethics in Internet. *J. Interdiscip. Stud.* **2020**, *32*, 179–192.
20. Wood, N. Protecting intellectual property on the Internet. Experience and strategies of trade mark owners in a time of chance. *Int. Rev. Law Comput. Technol.* **1999**, *13*, 21–28.
21. WIPO—World Intellectual Property Organization Magazine. Available online: https://www.wipo.int/wipo_magazine/en/ (accessed on 21 October 2023).
22. Ayyar, R. New Technologies Unleash Creative Destruction. In *The WIPO Internet Treaties at 25*; Springer: Cham, Switzerland, 2023; pp. 99–122.
23. Harnowo, T. Law as Technological Control of the Infringement of Intellectual Property Rights in the Digital Era. *Corp. Trade Law Rev.* **2022**, *2*, 65–79.
24. Atanasova, I. Copyright infringement in digital environment. *Econ. Law* **2019**, *1*, 13–22.
25. Tanielian, A.R.; Kampan, P. Saving online copyright: Virtual markets need real intervention. *J. World Intellect. Prop.* **2019**, *22*, 375–395.
26. Karahalios, H. Appraisal of a Ship's Cybersecurity efficiency: The case of piracy. *J. Transp. Secur.* **2020**, *13*, 179–201.
27. Hristov, K. Artificial intelligence and the copyright survey. *J. Sci. Policy Gov.* **2020**, *16*, 1–18.
28. Mateus, A.M. Copyright Violation on the Internet: Extent and Approaches to Detection and Deterrence. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2011.

29. Agrawal, S.; Sureka, A. Copyright infringement detection of music videos on YouTube by mining video and uploader meta-data. In Proceedings of the International Conference on Big Data Analytics, Mysore, India, 16–18 December 2013; Springer: Cham, Switzerland, 2013; pp. 48–67.
30. Omar, N.A.; Zakuan, Z.Z.M.; Saian, R. Software piracy detection model using ant colony optimization algorithm. *J. Phys. Conf. Ser.* **2017**, *855*, 012031.
31. Gray, J.E.; Suzor, N.P. Playing with machines: Using machine learning to understand automated copyright enforcement at scale. *Big Data Soc.* **2020**, *7*, 2053951720919963.
32. Stolikj, M.; Jarnikov, D.; Wajs, A. Artificial intelligence for detecting media piracy. *SMPTE Motion Imaging J.* **2018**, *127*, 22–27.
33. Jilcha, L.A.; Kwak, J. Machine Learning-Based Advertisement Banner Identification Technique for Effective Piracy Website Detection Process. *CMC-Comput. Mater. Contin.* **2022**, *71*, 2883–2899.
34. Zhang, D.Y.; Badilla, J.; Tong, H.; Wang, D. An end-to-end scalable copyright detection system for online video sharing platforms. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 626–629.
35. Zhang, D.Y.; Li, Q.; Tong, H.; Badilla, J.; Zhang, Y.; Wang, D. Crowdsourcing-based copyright infringement detection in live video streams. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) Barcelona, Spain, 28–31 August 2018; pp. 367–374.
36. Acheampong, F.A.; Nunoo-Mensah, H.; Chen, W. Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artif. Intell. Rev.* **2021**, *54*, 5789–5829.
37. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Int. J. Surg.* **2021**, *88*, 105906.
38. EBSCOhost: Academic Search. Available online: https://www.ebsco.com/ (accessed on 21 October 2023).
39. Taylor and Francis Online Homepage. Available online: https://www.tandfonline.com/ (accessed on 21 October 2023).
40. SpringerLink: Home. Available online: https://link.springer.com/ (accessed on 21 October 2023).
41. Elsevier Wordmark. Available online: https://www.elsevier.com/search-results (accessed on 21 October 2023).
42. Oxford Academic. Available online: https://academic.oup.com/journals (accessed on 21 October 2023).
43. Wiley Online Library. Available online: https://onlinelibrary.wiley.com/ (accessed on 21 October 2023).
44. Scopus. Available online: https://www.scopus.com/home.uri (accessed on 21 October 2023).
45. IEEE Xplore. Available online: https://ieeexplore.ieee.org/Xplore/home.jsp (accessed on 21 October 2023).
46. Association for Computing Machinery. Available online: https://www.acm.org/MDPI (accessed on 21 October 2023).
47. MDPI—Publisher of Open Access Journals. Available online: https://www.mdpi.com/ (accessed on 21 October 2023).
48. Indrawan, A.; Stevens, G.; Brianto, G.M.; Gaol, F.L.; Oktavia, T. Legal protection of copyright on copyrighted content downloaded through the internet. In Proceedings of the 2020 5th International Conference on Intelligent Information Technology, Hanoi, Vietnam, 19–22 February 2020; pp. 97–101.
49. Stuckey, K.D. *Internet and Online Law*; Law Journal Press: New York, NY, USA, 2023.
50. Hartmann, I.A. A new framework for online content moderation. *Comput. Law Secur. Rev.* **2020**, *36*, 105376.
51. Quintais, J.P.; De Gregorio, G.; Magalhães, J.C. How platforms govern users' copyright-protected content: Exploring the power of private ordering and its implications. *Comput. Law Secur. Rev.* **2023**, *48*, 105792.
52. Litman, J. Revising copyright law for the information age. In *The Internet and Telecommunications Policy*; Routledge: Abingdon, UK, 2020; pp. 271–296.
53. Park, C.; Kim, S.; Wang, T. Multimedia copyright protection on the web-issues and suggestions. In Proceedings of the 2012 IEEE International Symposium on Multimedia, Irvine, CA, USA, 10–12 December 2012; pp. 274–277.
54. Ray, A.; Roy, S. Recent trends in image watermarking techniques for copyright protection: A survey. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 249–270.
55. Megías, D.; Kuribayashi, M.; Qureshi, A. Survey on decentralized fingerprinting solutions: Copyright protection through piracy tracing. *Computers* **2020**, *9*, 26.
56. Warren, M. Server Authentication and its Role in Controlling Access to Copyrighted Works in Software and Video Games. SSRN: Amsterdam, The Netherlands, 2017.
57. Jin, X.; Dang, F.; Fu, Q.A.; Li, L.; Peng, G.; Chen, X.; Liu, K.; Liu, Y. StreamingTag: A scalable piracy tracking solution for mobile streaming services. In Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, Sydney, Australia, 17–21 October 2022; pp. 596–608.
58. Li, Y.; Wang, J. Robust content fingerprinting algorithm based on invariant and hierarchical generative model. *Digit. Signal Process.* **2019**, *85*, 41–53.
59. Chikada, A.; Gupta, A. 14 Online brand protection. In *Handbook of Research on Counterfeiting and Illicit Trade;* Elgar: Northampton, MA, USA, 2017; p. 340.
60. Gupta, K.; Saxena, A. Traversing the Digital Intellectual Property Realm on Social Media: An Abyss of Exploitation. *Nauls Law J.* **2020**, *15*, 80.
61. Sinhal, R.; Jain, D.K.; Ansari, I.A. Machine learning based blind color image watermarking scheme for copyright protection. *Pattern Recognit. Lett.* **2021**, *145*, 171–177.

62. Galli, F.; Loreggia, A.; Sartor, G. The Regulation of Content Moderation. In Proceedings of the International Conference on the Legal Challenges of the Fourth Industrial Revolution, Lisbon, Portugal, 5–6 May 2022; Springer: Cham, Switzerland, 2022; pp. 63–87.
63. Chen, X.; Qu, X.; Qian, Y.; Zhang, Y. Music Recognition Using Blockchain Technology and Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 7025338.
64. Marsoof, A. 'Notice and takedown': A copyright perspective. *Queen Mary J. Intellect. Prop.* **2015**, *5*, 183–205.
65. Metalitz, S.J. Implementation of the DMCA: The Practical Experience. International Intellectual Property Law and Policy **2002**, *7*, 1.
66. Urban, J.M.; Karaganis, J.; Schofield, B.L. Notice and takedown: Online service provider and rightsholder accounts of everyday practice. *J. Copyr. Soc. USA* **2017**, *64*, 371.
67. Szwajdler, P. Limitations of the Freedom of Hyperlinking in the Fields of Copyright Law, Trademark Law and Unfair Competition Law: Is Case-by-case Approach Sufficient? *Comput. Law Secur. Rev.* **2022**, *45*, 105692.
68. Frosio, G.; Husovec, M. Accountability and responsibility of online intermediaries. In*The Oxford Handbook of Online Intermediary Liability*; Oxford University Press: Oxford, UK, 2020.
69. Zhang, D.Y.; Song, L.; Li, Q.; Zhang, Y.; Wang, D. Streamguard: A bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 901–910.
70. Roja, G.; Kakarla, A.; Jacob, T.P. Cyber Patrolling using Machine Learning. In Proceedings of the 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), IEEE, 7–9 April 2022; pp. 75–78.
71. Alawad, A.A. Exploring Google Reverse Image Search to Detect Visual Plagiarism in Interior Design. *J. High. Educ. Theory Pract.* **2021**, *21*, 198–208.
72. Hwang, J.; Kim, J.; Chi, S.; Seo, J. Development of training image database using web crawling for vision-based site monitoring. *Autom. Constr.* **2022**, *135*, 104141.
73. Georgoulias, D.; Pedersen, J.M.; Falch, M.; Vasilomanolakis, E. Botnet business models, takedown attempts, and the darkweb market: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–39.
74. Kim, D.; Heo, S.; Kang, J.; Kang, H.; Lee, S. A photo identification framework to prevent copyright infringement with manipulations. *Appl. Sci.* **2021**, *11*, 9194.
75. Noah, N.; Tayachew, A.; Ryan, S.; Das, S. PhisherCop: Developing an NLP-Based Automated Tool for Phishing Detection. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Atlanta, GA, USA, 10–14 October 2022; SAGE Publications: Los Angeles, CA, USA, 2022; Volume 66, pp. 2093–2097.
76. Google Images. Available online: https://images.google.com/ (accessed on 21 October 2023).
77. TinEye Reverse Image Search. Available online: https://tineye.com/ (accessed on 21 October 2023).
78. Yandex. Available online: https://yandex.com/ (accessed on 21 October 2023).
79. Schröer, C.; Kruse, F.; Gómez, J.M. A systematic literature review on applying CRISP-DM process model. *Procedia Comput. Sci.* **2021**, *181*, 526–534.
80. Rupapara, V.; Narra, M.; Gonda, N.K.; Thipparthy, K. Relevant data node extraction: A web data extraction method for non contagious data. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), IEEE, Coimbatore, India, 10–12 June 2020; pp. 500–505.
81. BrightEdge—Enterprise SEO Platform. Available online: https://www.brightedge.com/ (accessed on 21 October 2023).
82. Golden Tomato Awards: Best Movies & TV of 2022. Available online: https://editorial.rottentomatoes.com/rt-hub/golden-tomato-awards-2022/ (accessed on 21 October 2023).
83. The 50 Best Albums of 2022 Staff List Billboard. Available online: https://www.billboard.com/lists/best-albums-2022/ (accessed on 21 October 2023).
84. Most Popular Apps. Available online: https://www.microsoft.com/en-us/store/most-popular/apps/pc (accessed on 21 October 2023).
85. The 100 Must-Read Books of 2022. Available online: https://time.com/collection/must-read-books-2022/ (accessed on 21 October 2023).
86. Sawant, K.; Tiwari, R.; Vyas, S.; Sharma, P.; Anand, A.; Soni, S. Implementation of selenium automation & report generation using selenium web driver & ATF. In Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 19–20 February 2021; pp. 1–6.
87. Serach Engine Journal. Available online: https://www.searchenginejournal.com/seo/meet-search-engines/ (accessed on 21 October 2023).
88. Yahoo Search. Available online: https://search.yahoo.com/ (accessed on 21 October 2023).
89. Bing. Available online: https://www.bing.com/ (accessed on 21 October 2023).
90. DuckDuckGo — Privacy, simplified. Available online: https://duckduckgo.com/ (accessed on 21 October 2023).
91. Most Visited Websites. Available online: https://trends.netcraft.com/topsites (accessed on 21 October 2023).
92. Carpineto, C.; Re, D.L.; Romano, G. Using Information Retrieval to Evaluate Trustworthiness Assessment of Eshops. In Proceedings of the IIR, Lugano, Switzerland, 5–7 June 2017; pp. 1–8.

93. Zhu, S.; Zhang, Z.; Yang, L.; Song, L.; Wang, G. Benchmarking label dynamics of virustotal engines. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual, 9–13 November 2020; pp. 2081–2083.
94. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Proceedings of naacL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, p. 2.
95. Geetha, M.; Renuka, D.K. Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *Int. J. Intell. Netw.* **2021**, *2*, 64–69.
96. Huo, H.; Iwaihara, M. Utilizing BERT pretrained models with various fine-tune methods for subjectivity detection. In Proceedings of the Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, 18–20 September 2020; Springer: Cham, Switzerland 2020; pp. 270–284.
97. Gelenbe, E.; Yin, Y. Deep learning with dense random neural networks. In Proceedings of the Man-Machine Interactions 5: 5th International Conference on Man-Machine Interactions, ICMMI 2017 Kraków, Poland, 3–6 October 2017; Springer: Cham, Switzerland, 2018; pp. 3–18.
98. Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, L.K.; Olivares-Mercado, J.; Portillo-Portilo, J.; Avalos, J.G.; García Villalba, L.J. Detecting Cryptojacking Web Threats: An Approach with Autoencoders and Deep Dense Neural Networks. *Appl. Sci.* **2022**, *12*, 3234.
99. Abubakar, H.D.; Umar, M.; Bakale, M.A. Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. *SLU J. Sci. Technol.* **2022**, *4*, 27–33.
100. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150.
101. Nguyen, T.T.H.; Jatowt, A.; Coustaty, M.; Doucet, A. Survey of post-OCR processing approaches. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–37.
102. Al-Saqqa, S.; Awajan, A. The use of word2vec model in sentiment analysis: A survey. In Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control, Cairo, Egypt, 14–16 December 2019; pp. 39–43.
103. Srivastava, R.; Bharti, P.; Verma, P. A review on multipolarity in sentiment analysis. *Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces*; Springer: Singapore, 2022; pp. 163–172.
104. Czarnecki, W.M.; Tabor, J. Multithreshold entropy linear classifier: Theory and applications. *Expert Syst. Appl.* **2015**, *42*, 5591–5606.
105. Basso, F.P.; Pillat, R.M.; Oliveira, T.C.; Del Fabro, M.D. Generative adaptation of model transformation assets: Experiences, lessons and drawbacks. In Proceedings of the 29th Annual ACM Symposium on Applied Computing, Gyeongju, Republic of Korea, 24–28 March 2014; pp. 1027–1034.
106. Shrestha, A.; Mahmood, A. Review of deep learning algorithms and architectures. *IEEE Access* **2019**, *7*, 53040–53065.