

Detecting Information Operations Using Machine Learning Algorithms [†]

Velizar Varbanov * and Tatiana Atanasova 

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Bl. 2, 1113 Sofia, Bulgaria; tatiana.atanasova@iict.bas.bg

* Correspondence: velizar.varbanov@iict.bas.bg

[†] Presented at the 15th International Scientific Conference TechSys 2026—Engineering, Technologies and Systems, Plovdiv, Bulgaria, 14–16 May 2026.

Abstract

The proliferation of fake news poses a significant threat to democratic institutions, public trust, and crisis management, particularly as it becomes a central tactic in coordinated information operations. This paper explores the application of machine learning (ML) algorithms in detecting and classifying disinformation as a means of supporting experts engaged in combating influence campaigns. We utilize a labelled English dataset, and a custom collected Bulgarian news dataset gathered using version 1 of the NewsData.io API, as indicated by the /api/1/ endpoint. After translating the Bulgarian content and generating synthetic fake news from real articles, we construct a multilingual training set. We evaluate the performance of three ML models demonstrating that advanced ML approaches can significantly enhance the identification of disinformation. These findings highlight the potential for machine learning to assist intelligence analysts, cybersecurity professionals, and policy makers in detecting and countering modern information operations at scale.

Keywords: fake news; disinformation; information operations (IO); machine learning (ML); support vector machines (SVM); multinomial naïve Bayes; BERT; text classification; cybersecurity

1. Introduction

The rise of social media and digital news platforms has amplified the spread of fake news, misinformation, and propaganda [1]. These influence campaigns have been used to sway elections, undermine public trust, and incite civil unrest. Addressing this threat requires robust systems for automatic detection of disinformation. Machine learning algorithms offer a powerful approach for analyzing linguistic patterns and classifying articles as fake or real [2,3].

Fake news is often characterized by emotional language, biased phrasing, or outright fabrication. Unlike traditional propaganda, today's disinformation campaigns are increasingly automated and targeted. Sophisticated actors use social bots and troll farms to manipulate discourse and destabilize societies. This highlights the importance of real-time detection tools.

Recent investigations into global influence operations have revealed how coordinated networks of fake websites—posing as legitimate regional news outlets—are used to disseminate politically motivated content in multiple languages. These networks amplify disinformation narratives aligned with strategic national interests, often targeting elections, agricultural industries, and military presence abroad. Tactics include fabricated press



Academic Editors: Nikola G. Shakev,
Valyo N. Nikolov, Nikolay R.
Kakanakov and Sevil A.
Ahmed-Shieva

Published: 4 August 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

releases, impersonation of legitimate brands, and artificial amplification through fake social media accounts.

Furthermore, advanced persistent threat (APT) groups and nation-state actors have weaponized disinformation to achieve strategic geopolitical objectives. Campaigns attributed to threat actors aligned with Russia, China, Iran, and North Korea have exploited major global events and election cycles to promote tailored narratives. Examples include:

- The use of fabricated press releases, fake petitions, and impersonated media outlets to manipulate narratives surrounding food security and Western corporations.
- Large-scale coordinated information operations by PRC-aligned infrastructure leveraging over 130 fake websites in multiple languages to promote state-aligned content.
- Social engineering and phishing attacks incorporating war and geopolitical lures, such as the Russia–Ukraine and Israel–Hamas conflicts.
- AI-enhanced IO campaigns that utilize generative models to craft narratives, impersonate personas, and localize propaganda.
- Hybrid disinformation strategies where domestic influencers and foreign states collaborate to spread content across Telegram, TikTok, YouTube, and Twitter/X.

These developments illustrate that fake news is not just a product of rogue individuals but often stems from structured and state-supported campaigns. Addressing such threats requires a holistic, interdisciplinary approach, combining technological tools with geopolitical awareness.

Research in fake news detection has explored both traditional ML methods [1] and deep learning approaches. The study [4] categorized detection into content-based and context-based methods. The research [5] compared linguistic feature engineering with probabilistic classifiers. More recently, transformer-based models like BERT [6] have achieved state-of-the-art results in text classification tasks. Additionally, multilingual models have enabled detection across languages [7]. Other studies have examined social network behaviour and metadata in classifying fake news [8]. For example, using propagation graphs and user credibility scores has proven effective in some cases [9]. However, text-based classifiers remain critical, especially when limited metadata is available [10,11].

2. Materials and Methods

2.1. Lexical Frequency Analysis

A comparative lexical analysis of the Bulgarian news dataset and the synthetically generated corpus of fake news reveals striking overlap in the most frequently used terms.

The top words in both datasets include references such as Europe, Euro, Bulgaria, USA, Politics, and Economy, alongside contextual terms like Referendum, GERB, Coalition, and War. These keywords are strongly linked to key socio-political developments affecting Bulgaria and the broader geopolitical landscape in 2025 (Figure 1). Notable topics reflected in the real news corpus include the anticipated replacement of the Bulgarian Lev with the Euro, a referendum proposed by the Bulgarian president on this issue, the ongoing Russia-Ukraine conflict, escalating tensions in Gaza, and global political shifts driven by the U.S. presidential elections and changes tied to the Trump administration.

Crucially, these same high-frequency words also appear prominently in the fake news dataset (Figure 2). This is not coincidental—fake news content often mimics the lexical environment of real news to capture public attention and lend an aura of legitimacy. By embedding disinformation narratives within familiar, emotionally charged, or politically sensitive topics, fake news can more effectively engage and manipulate readers.

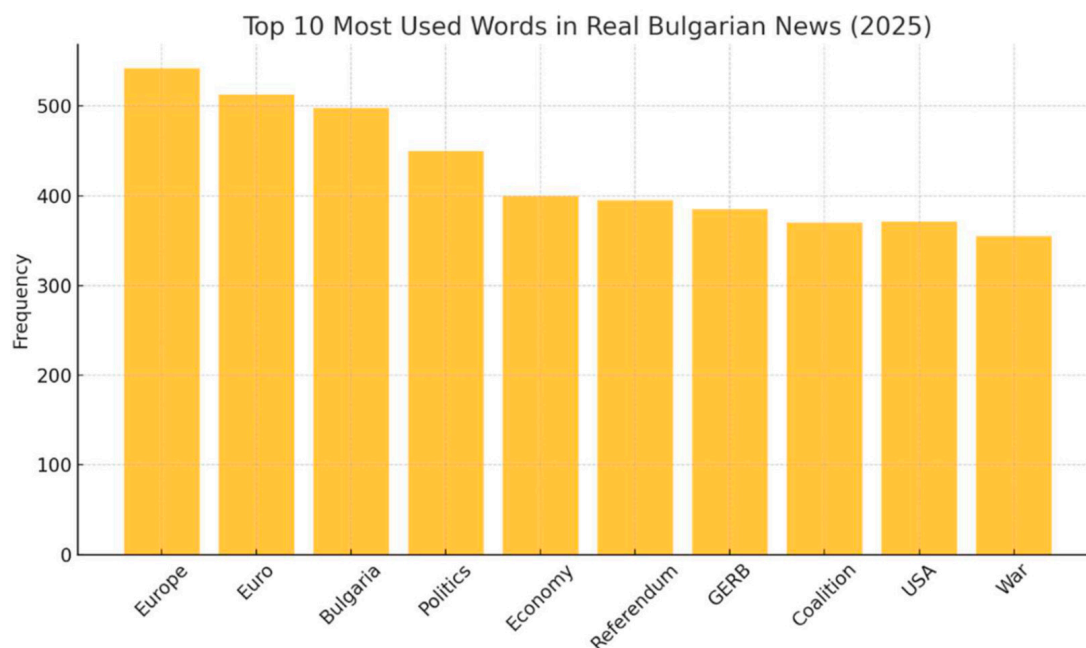


Figure 1. Top 10 most used words in real Bulgarian news.

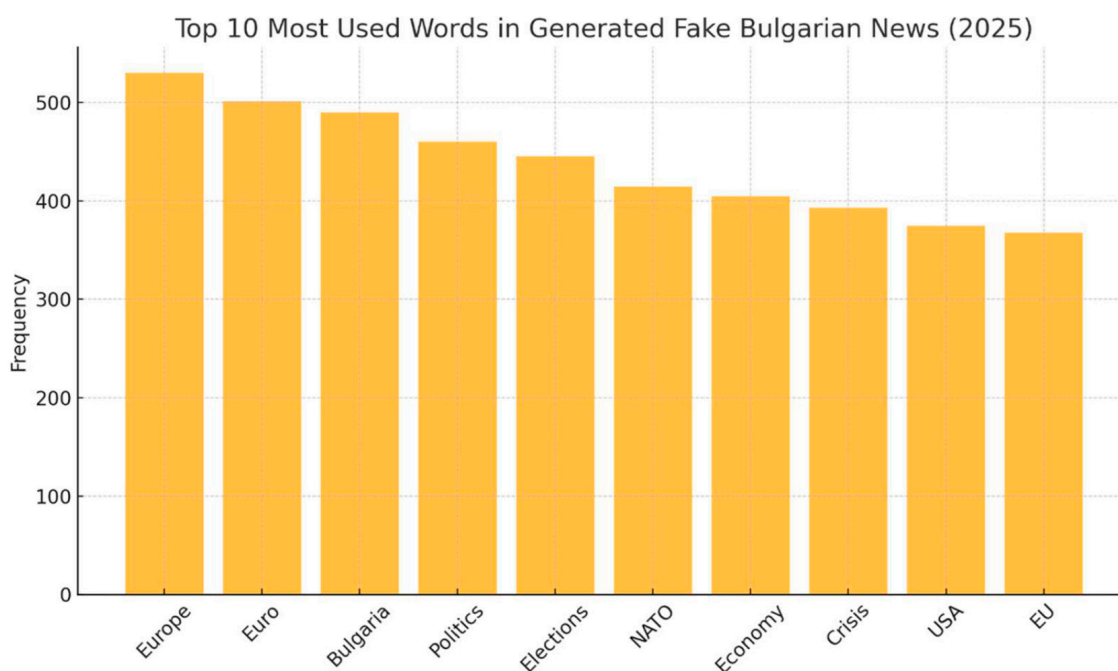


Figure 2. Top 10 most used words in generated fake Bulgarian news.

This lexical similarity between real and fake news illustrates the limitations of keyword-based approaches for disinformation detection. The presence of topical terms alone is insufficient to distinguish factual reporting from fabricated content, as both classes of texts are anchored in the same sociopolitical discourse. As such, superficial lexical analysis cannot provide reliable discrimination. This observation underscores the necessity of more sophisticated detection techniques, such as machine learning classifiers that leverage contextual, syntactic, and semantic cues. It reinforces our decision to adopt transformer-based models like BERT, which are capable of understanding deeper linguistic nuances beyond mere word frequencies.

2.2. Datasets

We used two primary datasets to train and evaluate machine learning models for fake news detection.

2.2.1. Fake and Real News Dataset (Kaggle)

The Fake and Real News Dataset, curated by Clément Bisailon and available on Kaggle, comprises two CSV files:

- Fake.csv: Contains 23,502 articles labelled as fake news.
- True.csv: Contains 21,417 articles labelled as real news.

Each entry in these files includes the following fields

- title: The headline of the news article.
- text: The full content of the article.
- subject: The category or topic of the news piece.
- date: The publication date of the article.

2.2.2. Bulgarian News Dataset

The dataset encompasses a diverse range of topics, including politics, world news, and government affairs. The fake news entries are sourced from unreliable or satirical outlets, while the real news entries originate from reputable news organizations. This dataset serves as a benchmark for binary text classification tasks in fake news detection.

To incorporate a multilingual perspective, a custom Bulgarian news dataset was compiled using the NewsData.io API. The data collection process involved:

- Language Filter: Articles were filtered to include only those written in Bulgarian.
- Country Filter: Only news sources based in Bulgaria were considered.
- Fields Extracted: For each article, the title, content, category, and publication date were retrieved.

The collected articles were saved in a CSV file with UTF-8 encoding to preserve Cyrillic characters. To prepare the data for analysis:

- Translation: The articles were translated into English using OpenAI's GPT-4 model to ensure consistency with the Kaggle dataset.
- Synthetic Fake News Generation: Utilizing GPT-4, synthetic fake news articles were generated by introducing subtle factual inaccuracies and sensational language into real news articles. This approach aimed to simulate realistic fake news content for training purposes.

The integration of this Bulgarian dataset enhances the model's ability to detect fake news across different languages and cultural contexts, addressing the challenges posed by multilingual disinformation campaigns.

2.3. Methodology

Our data preprocessing workflow was divided into two pivotal stages: initial preparation and thorough cleaning.

The preparation stage involved aggregating data from the aforementioned sources and segregating it to suit the training requirements of our ML models. The subsequent cleaning phase aimed at purging the data of inconsequential elements, such as common stop words, utilizing the NLTK Python library version 3.9.1, renowned for its natural language processing capabilities. This meticulous process unveiled a surprising observation: the most frequent terms across both datasets—regardless of their relevance—were strikingly similar, if not identical.

Such a similarity in keyword frequency across datasets casts doubts on the viability of keyword-based filtering for distinguishing fake from real news. This observation led to the pivotal decision to employ ML algorithms, which promise a more nuanced and effective approach to data classification. To this end, we introduced a “class” column within each dataset, assigning a ‘0’ for fake and a ‘1’ for real articles. This binary classification system is crucial for enabling the ML algorithms to discern the nature of the data during the training phase.

We employ a data preprocessing pipeline to facilitate the training and evaluation of our model. This pipeline encompasses two key steps: data splitting and TF-IDF vectorization.

Firstly, we partition our dataset into training and testing subsets using the `train_test_split` function from the `sklearn.model_selection` module. This process randomly separates our dataset into two portions: one for training, which constitutes 70% of the data, and another for testing, comprising 30% of the data. By ensuring a consistent random state (`random_state = 42`), we maintain reproducibility across experiments.

Subsequently, we leverage TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert textual data into numerical representations. We achieve this using the `TfidfVectorizer` from the `sklearn.feature_extraction.text` module. Initially, the vectorizer is fitted to the training data (`x_train`), which both learns the vocabulary and computes the TF-IDF weights. Then, both the training and testing textual data (`x_train` and `x_test`, respectively) are transformed into TF-IDF numerical vectors. This transformation ensures that our models—SVM, Multinomial Naïve Bayes, and BERT—can effectively interpret textual data, facilitating subsequent machine learning tasks.

- **Preprocessing:** Includes text normalization, tokenization, stop word removal, and TF-IDF vectorization. For BERT, raw text was tokenized using the pretrained tokenizer. Articles were lowercased, punctuation removed, and lemmatized.
- **Model Selection:** We evaluated three classifiers:
 - Support Vector Machine (SVM): Effective for high-dimensional feature spaces.
 - Multinomial Naïve Bayes (NB): Well-suited for text classification.
 - BERT (fine-tuned on the combined dataset): Pretrained transformer architecture with strong contextual understanding.
- **Training Setup:**
 - Train-test split (80/20) using stratified sampling.
 - Cross-validation with 5 folds.
 - Hyperparameter tuning using GridSearchCV (for SVM and NB).
 - Learning rate tuning and early stopping for BERT.
- **Evaluation Metrics:** Precision, Recall, F1-score, Accuracy, ROC-AUC.

3. Results

Our results demonstrate the robustness of machine learning algorithms in detecting and classifying fake news.

Among the models tested, BERT achieved the highest scores across all evaluation metrics (Table 1), indicating its superior capacity for understanding contextual language. This suggests that transformer-based architectures are particularly effective at identifying nuanced patterns of disinformation, including paraphrased or syntactically manipulated narratives.

These results underscore the potential of machine learning as a critical tool in the fight against information manipulation and hybrid threats. Hybrid threats encompass a blend of conventional and unconventional strategies, including cyberattacks, economic pressure, and disinformation campaigns. Information campaigns—deliberate efforts to influence,

mislead, or manipulate public perception—are increasingly being used as strategic weapons in geopolitical conflicts. Disinformation plays a central role in such campaigns, where it is weaponized to erode trust, destabilize societies, and undermine democratic processes.

Table 1. Performance results obtained from machine learning algorithms.

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Multinomial Naïve Bayes	0.89	0.88	0.90	0.89	0.91
SVM	0.91	0.92	0.90	0.91	0.93
BERT	0.95	0.96	0.94	0.95	0.97

In this context, machine learning can support intelligence analysts, cybersecurity professionals, and policy makers by automating the identification of disinformation patterns. Rather than relying solely on manual investigation or keyword matching, ML systems trained on multilingual and contextually enriched datasets can offer scalable, timely, and more accurate detection of malicious narratives. Integrating such tools into hybrid threat defence frameworks could enhance national resilience [12] and bolster the capacity to detect and counter influence operations at scale.

Our results (Figure 3) show that ML models can successfully classify fake news, including across languages. The inclusion of Bulgarian articles expanded the linguistic diversity of the dataset, and synthetic examples enhanced model generalization. The strong performance of BERT suggests that transformer-based architectures are especially well-suited for detecting subtle manipulations in text.

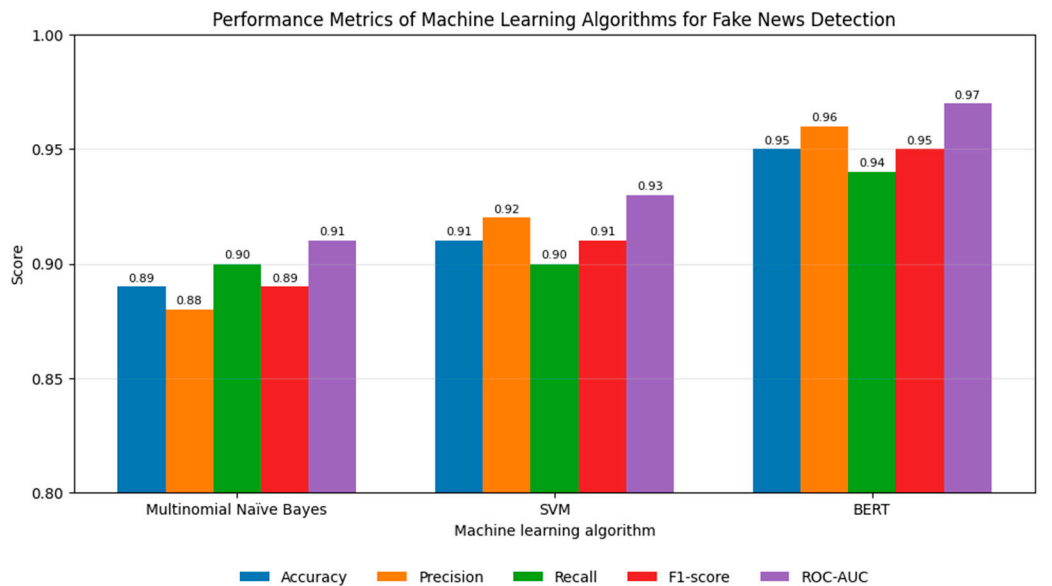


Figure 3. Performance metrics of machine learning algorithms for fake news detection.

Another notable observation is the high precision and recall scores of SVM, showing it remains a strong baseline for fake news detection. Naïve Bayes was computationally efficient and suitable for rapid prototyping, though it lagged behind in accuracy.

4. Discussion

The multilingual nature of our dataset adds a critical dimension, demonstrating how ML approaches can bridge linguistic gaps and uncover coordinated disinformation activities across regions. The ability to generalize across languages and topics is essential in an age where digital influence operations transcend borders and platforms.

Moreover, multilingual modelling is crucial for addressing regional disinformation. While English datasets dominate the research landscape, local-language fake news poses unique challenges. Our approach of translating and augmenting non-English content offers a scalable method for building robust classifiers.

The use of fabricated articles, impersonated organizations, and social media bots to seed disinformation adds further complexity to the challenge. For example, false petitions, cloned websites, and mirrored press releases have been used to target public sentiment around food security and international diplomacy. These operations can undermine trust in both digital and traditional media.

5. Conclusions

Fake news remains a persistent and evolving threat. By leveraging ML algorithms—especially modern language models like BERT—we can detect and counter disinformation with high accuracy. Future work includes expanding the multilingual dataset and incorporating context-based signals like metadata, propagation graphs, and user behaviour patterns.

- Collection of more regional and domain-specific fake news data.
- Use of datasets in Bulgarian without translating to English.
- Tests with other algorithms and neural networks
- Automated detection of fake domain networks and reused templates indicative of inauthentic news campaigns.

Furthermore, real-time deployment pipelines, integration with fact-checking systems, and adversarial robustness testing are critical directions for making fake news detection systems usable in production.

Author Contributions: Conceptualization, V.V. and T.A.; methodology, V.V.; software, V.V.; validation, V.V. and T.A.; investigation, V.V.; resources, V.V.; data curation, V.V.; writing—original draft preparation, V.V.; writing—review and editing, T.A.; visualization, V.V.; supervision, T.A.; project administration, T.A.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: The presented research is supported by the National Science Fund of Bulgaria, with Contract No. KII-06-H85/16.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

Abbreviations

The following abbreviations are used in this manuscript:

APT	Advanced Persistent Threat
PRC	People's Republic of China
IO	Information Operations
SVM	Support Vector Machine
NB	Multinomial Naïve Bayes

NLTK	Natural Language Toolkit
ML	Machine Learning
TF-IDF	Term Frequency–Inverse Document Frequency
BERT	Bidirectional Encoder Representations from Transformers
ROC-AUC	Receiver Operating Characteristic–Area Under Curve
AI	Artificial Intelligence
API	Application Programming Interface
CSV	Comma-Separated Values

References

1. Hu, B.; Mao, Z.; Zhang, Y. An overview of fake news detection: From a new perspective. *Fundam. Res.* **2025**, *5*, 332–346. [[CrossRef](#)] [[PubMed](#)]
2. Villela, H.F.; Corrêa, F.; de ANRibeiro, J.S.; Rabelo, A.; Carvalho, D.B.F. Fake news detection: A systematic literature review of machine learning algorithms and datasets. *J. Interact. Syst.* **2023**, *14*, 47–58. [[CrossRef](#)]
3. Khanam, Z.; Alwasel, B.N.; Sirafi, H.; Rashid, M. Fake News Detection Using Machine Learning Approaches. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2021; Volume 1099, p. 012040.
4. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [[CrossRef](#)]
5. Conroy, N.J.; Rubin, V.L.; Chen, Y. Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *25*, 1–4. [[CrossRef](#)]
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [[CrossRef](#)]
7. Hashmi, E.; Yildirim Yayilgan, S.; Hameed, I.A.; Mudassar Yamin, M.; Ullah, M.; Abomhara, M. Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration. *IEEE Access* **2024**, *12*, 121507–121537. [[CrossRef](#)]
8. Barbieri, F.; Camacho-Collados, J.; Neves, L.; Espinosa-Anke, L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1644–1650.
9. Bozarth, L.; Budak, C. Market Forces: Quantifying the Role of Top Credible Ad Servers in the Fake News Ecosystem. *ICWSM* **2021**, *15*, 83–94. [[CrossRef](#)]
10. Mihaylov, T.; Georgiev, G.; Nakov, P. Finding opinion manipulation trolls in news community forums. In Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL 2015), Beijing, China, 30–31 July 2015.
11. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. In *ACM Computing Surveys (CSUR)*; Association for Computing Machinery: New York, NY, USA, 2020; Volume 53, pp. 1–40.
12. Varbanov, V. Advancing Cyber Threat Intelligence through Machine Learning Algorithms. In Proceedings of the AICCONF '24: Cognitive Models and Artificial Intelligence Conference, İstanbul, Türkiye, 25–26 May 2024.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.