Jackson Hutton

3302

2/05/2025

Using AI to Fight AI: Detecting Fake News and Misinformation Using Machine Learning

Introduction

The internet has reshaped the way we share, discover, and discuss information, bringing both benefits and challenges. While real-time updates and social media posts can enrich conversations, they also make it harder to spot the difference between factual content and deliberate misinformation - or “fake news.” This issue has become even more urgent as online platforms increasingly serve as primary news sources, and posts with questionable credibility can go viral in moments. When false narratives spread, they can sway public opinion, create distrust in institutions, and even affect significant events like elections or health initiatives.

In an MIT study analyzing the proliferation of Fake News on social media, Vosoughi, Roy, and Aral (2018) found that false stories on Twitter are about 70% more likely to be retweeted than true ones, and they reach their first 1,500 viewers up to six times faster. Such alarming statistics indicate that misinformation easily outpaces efforts to identify and counteract it, placing a significant burden on individuals, media organizations, and policymakers. As these misleading narratives proliferate, they erode public trust and distort public discourse, creating a critical need for more sophisticated and scalable detection strategies.

Research on using machine learning to detect fake news is especially crucial because these algorithms can flag misleading stories quickly and at a scale that human reviewers alone cannot match. As misinformation tactics grow more sophisticated - through tactics like deepfakes or AI-generated text - machine learning offers a proactive way to identify emerging threats. With recent large-scale reductions in human moderation across big social media platforms, it is becoming more essential every day to have platforms that are capable of flagging misinformation automatically. In this literature review, I will be exploring how machine learning can be used as a tool to identify, classify, and respond to fake news and how researchers can help improve the reliability of digital media and protect the integrity of public discourse. These discussions will be accomplished through thorough review, critique, and synthesization of the literature reviewed.

Method

To compile the sources for this literature review, I began at Northeastern's A–Z databases list, where I filtered by the computer science category. From there, I selected Web of Science (All Databases) because it offers peer-reviewed, high-quality materials from a range of journals and conference proceedings. Using a search for topics both 'artificial intelligence' and 'fake news,' I initially found over two thousand results. To make the scope more focused, I limited the publication date range to 2020 through the present, ensuring that I captured only the most recent research.

Next, I further refined the results by selecting only 'highly cited papers' that were also open access. This helped to ensure that the final sources would not only be credible and impactful but also fully accessible for review. After these filters, I was left with 12 relevant articles. They were all relevant to the discussion I was trying to create, so I decided to include them all. No other keyword filters or exclusion criteria were applied at this point. These 12 works now serve as the core set of references for my literature review on applying machine learning techniques to detect fake news.

The resulting set featured articles from leading computer science and information technology journals, such as *Information Processing & Management*, *Future Generation Computer Systems*, *IEEE Transactions on Knowledge and Data Engineering*, and *Multimedia*

Tools and Applications. Focusing on peer-reviewed journals with strong citation rates and recent publication dates confirms that the sources are both trustworthy and up to date, providing a solid base for the discussion and analysis to follow. Additionally, the variety in journals helps to cast a wider net on the literature to be reviewed, which is important when discussing the topic as a whole.

Findings- Overview

In reviewing these 12 studies, it became clear that researchers generally take one of two main approaches when building fake news detection models; the first relies on *classical ensemble* methods- combining traditional machine learning algorithms like SVM or Random Forest- often with specifically crafted text features (Ahmad et al., 2020; Ozbay & Alatas, 2020; Hakak et al., 2021). These classical ensemble-type approaches focus on patterns in the text itself and usually follow a more traditional ML pipeline of feature extraction followed by classification. The second approach, discussed by the other papers, leverages *deep learning frameworks,* especially transformer-based architectures such as BERT, to capture richer contextual cues (Guo et al., 2024; Kaliyar et al., 2021; Qin & Zhang, 2024; Zhang et al., 2023). Some of these studies remain *text only,* while others expand into *multimodal* territory by incorporating images or social context into the detection process (Jing et al., 2023; Luvembe et al., 2024; Ma et al., 2023). Conjunctively, while many efforts focus exclusively on the binary task of real versus fake classification, a few researchers opt for *multi-task* or specialized methods, like handling multilingual content (Guo et al., 2024) or coupling fake news classification with stance and sentiment analysis (Liao et al., 2022). Although most of the cited works focus on quantitative measures (e.g., accuracy, precision, recall) to evaluate their machine learning models, several also include qualitative assessments, such as user feedback or content-based analysis, to better understand how their models perform in real-world settings. In total, 12 articles met the final criteria for inclusion in this review, all published between 2020 and 2024. Specifically, 2 were published in 2020, 3 in 2021, 1 in 2022, 3 in 2023, and 3 in 2024. All of these studies appeared in peer-reviewed journals, encompassing fields such as computer science and information technology. To facilitate comparison and identify overarching trends and gaps, the main body of this literature review is organized according to the methodological approaches described in the articles, moving from classical ensemble models to deep learning frameworks and finally addressing specialized or multimodal strategies, in which I will synthesize the findings of these 12 articles.

Classical Ensemble Methods

Classical ensemble approaches, as demonstrated by Ozbay and Alatas (2020) and Ahmad et al. (2020) frequently combine multiple traditional machine learning algorithms such as Support Vector Machines, Decision Trees, or Random Forest to capture different facets of textual patterns. In the article by Hakak et al. (2021), other, more customized linguistic features were integrated into ensemble pipelines to improve classification breadth. In comparing these studies, researchers generally looked to balance two important factors of applied ML techniques-interpretability and efficiency. In balancing these two, conventional algorithms are able to remain relevant, especially when appropriately tuned and combined. However, these ensemble-based approaches often rely on feature engineering, are more unpredictable, and may lack the adaptability to handle the increasing complexity of online fake news. Additionally, these approaches are somewhat outdated compared to the next two approaches we will discuss in this literature review, as the application of these methods occurred mostly in 2020 and 2021 according to the papers I reviewed.

Deep Learning Frameworks

Conversely, a separate set of researchers employs deep learning methods, particularly transformer-based architectures. Guo et al. (2024) and Qin and Zhang (2024) explore how language models like BERT can capture nuanced linguistic cues that simpler models might overlook. Likewise, Kaliyar et al. (2021) implements a fine-tuned BERT workflow to derive richer contextual insights without extensive manual feature extraction. Zhang et al. (2023) also experiments with convolutional and recurrent neural networks in addition to transformers, illustrating the broader trend of leveraging advanced neural layers to detect patterns in text more dynamically. Together, these studies indicate that, because of recent advancements in the technology, deep learning architectures not only reduce the reliance on manual feature engineering but also adapt more readily to shifting linguistic norms in fake news content.

Specialized and Multimodal Strategies

A growing body of research broadens fake news detection beyond standard text-only inputs. According to Sahoo and Gupta (2021), incorporating social engagement signals, like user profiles, comment histories, or share patterns, can reveal deceptive behavior that pure text-based approaches might miss. Meanwhile, Luvembe et al. (2024) and Jing et al. (2023) use multimodal frameworks, integrating image or video analysis into detection pipelines, suggesting that blended data streams help identify evolving misinformation tactics. Similarly, Ma et al. (2023) proposes graph-based techniques to capture relationship structures among users and posts, whereas Liao et al. (2022) employs a multi-task model to address stance detection and sentiment analysis alongside fake news classification. In expanding the scope of input features and tasks, these specialized and multimodal systems are a good representation of the field’s shift in recent years toward holistic methods that tackle both textual and contextual aspects of misinformation. This shift is important for accurately classifying fake news, as user context is something that purely text input-based approaches struggle to understand.

Conclusion

All in all, these papers underline how machine learning is increasingly central to detecting fake news across digital platforms. Classical ensemble approaches emphasize interpretable pipelines and have demonstrated solid performance, especially when carefully tuned or enhanced by linguistically inspired features. However, their reliance on hand-engineered inputs may, unfortunately, limit their capacity to adapt rapidly as misinformation evolves. Deep learning frameworks, particularly those leveraging transformer-based architectures, offer more robust approaches by automatically learning complex linguistic cues and reducing the need for exhaustive feature engineering. At the same time, specialized and multimodal research that integrates visual, social, or user-based metadata is very exciting in an age when manipulated images and AI generated content can magnify the impact of false narratives.

The literature I reviewed suggests that the most effective detection models will likely merge the strengths of these disparate paradigms and are trending toward more unified approaches to tackle the problem effectively. As researchers expand their methods to handle multilingual and multimodal content, the field is moving closer to tools that can contend with the sheer volume of online misinformation. Continuing to progress in this field, however, will demand ongoing interdisciplinary collaboration, along with policy-level and platform-driven initiatives to incorporate advanced AI systems responsibly, and can overall be a net positive on user experience on the internet as a whole.

References

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity, 2020,* Article 8885861. https://doi.org/10.1155/2020/8885861

Guo, Z. W., Zhang, Q., Ding, F., Zhu, X. G., & Yu, K. P. (2024). A novel fake news detection model for context of mixed languages through multiscale transformer. *IEEE Transactions on Computational Social Systems*, *11*(4), 5079–5089. https://doi.org/10.1109/TCSS.2023.3298480

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, *117*, 47–58. https://doi.org/10.1016/j.future.2020.11.022

Jing, J., Wu, H. C., Sun, J., Fang, X. C., & Zhang, H. X. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management, 60*(1), Article 103120. https://doi.org/10.1016/j.ipm.2022.103120

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

Liao, Q., Chai, H. Y., Han, H., Zhang, X., Wang, X., Xia, W., & Ding, Y. (2022). An integrated multi-task model for fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, *34*(11), 5154–5165. https://doi.org/10.1109/TKDE.2021.3054993

Luvembe, A. M., Li, W. M., Li, S. H., Liu, F. F., & Wu, X. (2024). Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management, 61*(3), Article 103653. https://doi.org/10.1016/j.ipm.2024.103653

Ma, X. X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., & Akoglu, L. (2023). A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering, 35*(12), 12012–12038. https://doi.org/10.1109/TKDE.2021.3118815

Ozbay, F. A., & Alatas, B. (2020). *Fake news detection within online social media using supervised artificial intelligence algorithms*. *Physica A: Statistical Mechanics and Its Applications*, *540*, 123174. https://doi.org/10.1016/j.physa.2019.123174

Qin, S. M., & Zhang, M. L. (2024). *Boosting generalization of fine-tuning BERT for fake news detection*. *Information Processing & Management*, *61*(4), Article 103745. https://doi.org/10.1016/j.ipm.2024.103745

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing, 100*, Article 106983. https://doi.org/10.1016/j.asoc.2020.106983

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zhang, Q., Guo, Z. W., Zhu, Y. Y., Vijayakumar, P., Castiglione, A., & Gupta, B. B. (2023). A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters, 168*, 31–38. https://doi.org/10.1016/j.patrec.2023.02.026