

up stories, edited images and videos, misleading headlines, satire that is not actually factual, and a coordinated campaign of disinformation organized by state and non-state actors (Zhou and Zafarani, 2020). The impact of unregulated fake news is extensive: the COVID-19 pandemic, in its turn, was accompanied by a parallel infodemic where misleading information about health could beget misinformation about vaccination and the destruction of the efforts of interventions to support the health of the population worldwide (Cinelli et al., 2020). Likewise, the fake news has been cited to contribute to the heightened political polarization, violence, and loss of confidence in the democratic establishments (Vosoughi et al., 2018).

Conventional methods to counterfeit news, including manual verification by websites like Snopes, PolitiFact, and FactCheck.org, are not sufficient in dealing with the amount and rate of fake news circulating throughout social media (Thorne et al., 2018). The professionalism and correctness of human fact-checkers cannot be sufficient in scope; the amount of material published every day in social media is overwhelming; no agency can check it all. Moreover, the economic factors that support the social media advertising framework tend to focus more on engagement than the accuracy since sensational information, a feature that is often related to fake news, is more likely to attract user interaction (Lazer et al., 2018). Such structural issues have pushed researchers, technologists, and policymakers to contemplate artificial intelligence more and more as a scalable solution in the field of automated detection of fake news.

Artificial intelligence (AI), and machine learning (ML) specifically, has a potential for transforming the phenomenon of detecting and reducing fake news on social media. Using massive collections of annotated news stories, social media posts, and user behavior patterns, AI systems can be conditioned to identify linguistic signals, propagation features, and signs of credibility of the source that allow distinguishing between genuine news and fake news (Shu et al., 2019). Transformer-based language models, like BERT and GPT, which have shown exceptional precision in detecting deceptive framing, emotional manipulation, and inaccurate facts, have been demonstrated to be useful in natural language processing (Devlin et al., 2019). These models are capable of assessing thousands of articles in a single second and therefore are much more appropriate than human researchers to the demands of social media moderation in real-time.

Outside of the textual analysis, multimodal AI methods have broadened the use of fake news detection to encompass image, video, and metadata analysis. Convolutional neural networks (CNNs) enable the training of deepfake detection algorithms that can detect insidious artifacts of the manipulated media that humans are unable to detect (Tolosana et al., 2020). The machine learning models based on graphs can study the social network topology according to which the information spreads and differentiate between diffusion trends of verified information and misinformation depending on the velocity of reposting and credibility profile of key influential figures (Bian et al., 2020). In addition, knowledge graph methods take advantage of structured collections of known facts and automatically look them up to investigate claims against trusted sources, which allows more subtle, context-specific detection (Shi and Weninger, 2016).

Although these are promising developments, the introduction of AI-based systems of detecting fake news presents considerable technical, ethical, and social issues. Adversarial attacks have the vulnerability of attempting to manipulate AI models whereby bad players develop specific content to avoid detection algorithms (Zellers et al., 2019). AI detection systems are highly sensitive to the quality and representativeness of training data, which is usually biased towards the content of the English language and Western culture, leading to a decrease in performance in other linguistic and cultural environments (Hardalov et al., 2022). Ethically, automated content moderation poses great challenges in terms of the freedom of expression, algorithm bias, transparency and accountability (Wardle and Derakhshan, 2017).

The article is a review of the existing situation with AI-based fake news detectors on social media, it discusses approaches, datasets, performance indicators, and actual implementations. It also critically analyzes technical constraints and ethical considerations of automated detection mechanisms and reveals major directions of future studies. Having combined the findings of machine learning, natural language processing, social network analysis, and media studies, this review will add to a more refined perspective on how AI can be productively and conscientiously utilized to ensure the integrity of the information environment in the digital era.

2. Literature Review

2.1 The Fake News: The Notion and Taxonomies

There is a great deal of divergence in the academic literature on the specific definition of fake news, which has been used with a loose interpretation to refer to a heterogeneous set of information disorders. Wardle and Derakhshan (2017) suggested the basic typology that differentiates between misinformation (the false content shared without an intention to harm), disinformation (the false content created with intent to deceive), and malinformation (the true content shared with intent to harm). This tri-faceted differentiation has generally been incorporated in the future studies as a framework to more detailed methods of detection that puts into consideration not only veracity, but also intent and possible harm. Tandoc et al. (2018) also developed the typology of fake news, depending on the issues of facticity and degree of deception and identified six categories of fake news: news satire, news parody, fabrication, manipulation, advertising, and propaganda.

One of the earliest empirically-based studies of fake news, reported by Allcott and Gentzkow (2017), focused on the 2016 U.S. presidential election and reported the magnitude of exposure of fake news and the strong contribution of social media platforms to spreading fabricated information. They have shown that false stories were viewed an average of 760 million times during the election campaign, and consumption was concentrated among older, politically partisan users. These results highlighted the size of the problem and the necessity of automated detection solutions with the ability to perform at the social media scale. Shu et al. (2017) then formulated the detection of fake news as an official machine learning task, with initial benchmark datasets and taxonomies which have since become standard points of reference in the domain.

It has also been an issue of considerable academic interest as to the role of cognitive and behavioral factors in consumption and sharing of fake news. Pennycook and Rand (2019) have proved that the vulnerability to fake news does not depend solely on political partisanship or driven reasoning, but it can be greatly predicted by the lack of analytical processing and media literacy skills. The implications of this discovery on AI detection systems are significant because it is essential to consider the reality that fake news is frequently very plausible and emotionally appealing and is created to capitalize on cognitive biases that are close to any social media user. Psychological vulnerability and algorithmic amplification meet each other in this way, therefore, forming a complicated environment of detection that can hardly be covered by purely computational methods (Pennycook and Rand, 2021).

2.2 Machine Learning and Natural Language Processing

Initial computational solutions to detect fake news were mainly based on feature engineering with classical machine learning classifiers like Support Vector Machines (SVM), Naive Bayes and logistic regression. These techniques derived linguistic cues, such as patterns of punctuations, capital letters, emotional value, and readability scores, out of the news articles text and trained the classifiers to learn how to tell a genuine article and a forged one (Castillo et al., 2011). Although these methods provided significant benchmarks, the fact that they were based on features that were manually engineered restricted their scalability and effectiveness to novel types of fake news that utilized alternative linguistic resources.

Deep learning revolutionized the field and made it possible to learn hierarchical representations directly on raw text and end to end. Recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks were found to have significantly increased performance when compared to classical approaches because they were able to work out sequential dependencies in text (Wang, 2017). A new paradigm change was introduced with attention mechanisms and transformer architectures; BERT and its variations set state-of-the-art scores on fake news detection challenges at the expense of pre-training large corpora to construct rich and contextually aware representations of language (Devlin et al., 2019). Subsequently, it was shown that domain-adaptive fine-tuning of BERT on fake news could achieve a detection accuracy of over 90 per cent on known data sets, a significant improvement over previous methods (Kula et al., 2021).

Stance detection- deciding whether a particular piece of writing will be in support, or in refutation, or neutral to a particular claim or not- became a valuable related task to fake news detection. In 2017, the Fake News Challenge (FNC-1), which offered a standardized stance hands-on assessment, was launched, and it resulted in a lot of research being conducted (Thorne et al., 2018). Misinformation detection NLP toolkit has also been extended by claim verification methods that use systematic knowledge bases and external indicators to determine the truthfulness of particular claims. FEVER dataset, an 185,000 claim-evidence pairs dataset, has become a universal reference point of testing claim verification systems (Thorne et al., 2018). These solutions indicate a wider trend in the area towards evidence-based, as opposed to simply stylistic, technique of fake news detection.

2.3 Propagation-Based Detection and Social Network Analysis

Another strand of literature has concentrated not on what is contained in the news stories themselves, but on how the social network forms and propagation dynamics of information dissemination. The landmark research study by Vosoughi et al. (2018), which took place in Science, revealed that the spreading of the false news is much more rapid, widespread, and profound than the spreading of true news on Twitter, despite the novelty and emotional involvement of the information. This biased propagation structure offers a strong signal to detection algorithms; graph-based models which examine the network topology of information propagation have shown strong classification results without necessarily depending on content analysis.

Bian et al. (2020) presented a bi-directional graph convolutional network (Bi-GCN), simultaneously modeling both top-down propagation and bottom-up dispersion patterns in rumor diffusion trees, and reaching the state of the art on the PHEME and Twitter15/16 datasets. Graph attention networks (GATs) have also made this method more sophisticated by allowing the models to assign dissimilar weights to the edges and nodes of the propagation graph, representing the different credibility and influence of the various users in the diffusion network (Shu et al., 2019). These graph-based models have been found to incorporate user credibility signals such as account age, number of followers, past posting behavior, and network centrality, which have been shown to strongly augment the level of detection, especially at an early stage of detecting an emerging false narrative (Ruchansky et al., 2017).

2.4 Multimodal Detection and Deepfake Identification

Since the deployment of fake news has been much aided by visual and audiovisual information, scholars have come up with multimodal detection models that can process the text, image, and video simultaneously. Wang et al. (2018) showed that a combination of visual and textual features obtained with the help of CNNs and transformer models produced significantly higher detection rates as compared to when using unimodal features to deal with news articles that include images to mislead but no false text. The Multimodal Variational AutoEncoder (MVAE) model which is presented by Khattar et al. (2019)

learns both text and image representations simultaneously, and is sensitive to cross-modal inconsistencies inherent to fake news articles.

With the advent of the so-called deepfakes, or hyper-realistic synthetic media created by generative adversarial networks (GANs), there are now new problems in the field of fake news detection that go beyond the conventional NLP. Tolosana et al. (2020) have given an in-depth overview of deepfake generation and detection methods, recording an arms race in technologies where increasingly more advanced generation algorithms have been counteracted by increasingly sophisticated detection models. Detection models based on CNNs, which exhibit frequency-domain features of synthetic images, have been shown to perform well; but these models are susceptible to compression and post-processing algorithms, which are routinely used on the content of social media. FaceForensics++ has become the new standard of deepfake detecting research (Rossler et al., 2019).

3. METHODOLOGY

3.1 Research Design

This research paper will use systematic literature review as its research design and take the PRISMA (Preferred Reporting Items to Systematic Reviews and Meta-Analyses) framework as its guiding tool to synthesize the existing literature on the topic of AI-based fake news detection. The systematic review method was chosen as it offers a rigorous, transparent, and reproducible way of locating, appraising, and synthesizing the results of the body of the available literature (Liberati et al., 2009). The research questions that would inform this review are: (1) What AI methods have been used to detect fake news on social media and what are their performance results? (2) What are the popular datasets and evaluation benchmarks in the field? (3) What are the major technical and ethical drawbacks of existing AI detection methods? (4) What does the evidence tell us does research and practice have to do to go?

3.2 Search Strategy and Inclusion Criteria

The search strategy included articles published after 2008 and with a focus on the health topic prior to 2008. Four large scholarly databases, namely IEEE Xplore, ACM Digital Library, Google Scholar, and Web of Science, were searched through in full. The search was carried out in January 2025, and entailed the period between January 2014 and December 2024. The search query was made out of terms describing the phenomenon of interest (Fake news, misinformation, disinformation, rumor detection) along with terms describing the methods of analysis of interest (machine learning, deep learning, artificial intelligence, natural language processing, neural network, graph neural network). Only peer-reviewed journal articles and conference proceedings as well as book chapters in English language were searched. The studies were included: (a) they postulated, assessed, or both, AI-based methods of fake news or misinformation detection; (b) they were empirically based, and their results were quantitative, on established benchmarks; (c) they studied the specific context of social media.

3.3 Analysis of Data and Extraction

Two reviewers then independently screened titles and abstracts to determine their suitability to the inclusion criteria after initial database searches. Article full texts of the potentially eligible articles were then accessed and evaluated accordingly. Data were picked out with the help of a standardized extraction form, which included: author(s) and year of publicity; AI methodology applied; dataset(s) applied; measurements of evaluation reported; main findings; and limitations found. Themes were synthesized on extracted data and studies were grouped by the main AI paradigm used: (a) classical machine learning, (b) deep learning and NLP, (c) graph-based and propagation models, (d) multimodal detection and (e) knowledge graph and claim verification methods. Accuracy, F1-score, precision and recall were tabulated to facilitate cross-study comparison where methodological comparable could be done.

4. Data Analysis and Findings

4.1 Comparison of AI Approaches on Performance

The systematic review has found 87 original studies that fit the inclusion criteria and covered the entire spectrum of AI paradigms used to detect fake news. Comparison of the reported measures of performance shows a significant trend in accuracy in detection in line with the improvement of the model architecture. The classical machine learning methods, SVM, Naive Bayes, and logistic regression models on handcrafted features, obtained average scores of 72-81% according to popular datasets such as the LIAR (Wang, 2017). Deep learning models based on LSTMs were better than these baselines with an accuracy score of 82-87. Models that are based on transformers, especially fine-tuned versions of BERT, RoBERTa, and XLNet, were steadily scoring the highest accuracy scores, and a few reported F1-scores above 92 percent on the LIAR, FakeNewsNet, and FEVER benchmarks (Kula et al., 2021).

Graph-based models showed especially good ability in rumor detection tasks which use propagation data. Bi-GCN model suggested by Bian et al. (2020) was capable of reaching 88.6% accuracy on the Twitter15 dataset and 89.1% on the Twitter16 dataset, which is significantly higher than content-only models, especially in case of early-stage detection tasks. Models that used a combination of either text or visual characteristics showed steady improvements over unimodal counterparts in datasets that used image-text pairs and the accuracy of the models increased by 3-7 percentage points. This work implies that one AI paradigm cannot be considered superior to the rest; instead, the best way to do something is task-specific and depends on the accessibility of propagation data, visual data, and the time limitations related to a detection scenario.

4.2 Dataset Analysis

The comparison of the datasets employed across studies showed that there is a high level of heterogeneity, which is a big impediment to cross-study comparability. The datasets that have been most frequently used are: LIAR (12,836 labeled statements of PolitiFact), FakeNewsNet (news content with social contexts of GossipCop and PolitiFact), PHEME (rumor threads of Twitter), BuzzFeed News (Facebook posts of the 2016 U.S. election), and FakeHealth (health-related misinformation). One of the most important conclusions made during the analysis of the dataset is that the vast majority of benchmark datasets consist of language materials of predominantly English language, and there are very few datasets of other languages, including Arabic, Hindi, Urdu, or Mandarin. This bias of the English language is a major barrier to the use of detection models trained on these datasets to the universal social media context (Hardalov et al., 2022).

4.3 Ethical and Bias Analysis

The content analysis of the studies examined showed that the issue of ethics is not considered in most of the research on AI fake news detection. Only a quarter of the analyzed articles directly covered the possible problems of algorithmic biases in their detection models; less than a fifth of them touched upon the consequences of false positives, which are real news but wrongly identified as fake, and their effects on freedom of expression. The analysis found three main types of bias in current systems of detecting them: (a) selection bias based on non-representative training samples; (b) confirmation bias due to labeling mechanisms that are based on human annotators whose political or cultural views are subject to change over time; and (c) temporal bias based on the concept drift so that the linguistic and stylistic features of fake news gradually change over time and cannot be tracked by the model that operates solely on the snapshot (Zhou and Zafarani, 2020). These biases are important implications to the practical use of automated detection systems on a large scale.

5. Discussion

The systematic review of AI-based solutions to fake news detection in social media proves that the development of technologies has achieved a high level of detection, but there are still issues at the technical, ethical, and functional levels. The paradigm comparison shows that classical methods of machine learning, like Support Vector Machines (SVM) and Naive Bayes, are capable of delivering sensible performance on small datasets, but due to their dependence on manually-crafted features, they do not scale or adapt to new fake news trends. Deep learning models, especially based on LSTM networks, have better accuracy as they enable the sequential dependency of textual data; but the true breakthrough is transformer-based models such as BERT, RoBERTa, and XLNet. The models are more successful in context-sensitive detection tasks and have shown the highest F1-scores on the existing benchmarks, which underscores their higher ability to process complex linguistic structure, emotion indicators, and subtle misleading information.

Graph-based propagation models offer a complementary approach as they look at the social pattern of information dissemination. The Bi-GCN and the Graph Attention Networks (GATs) have exhibited especially good performance in early detection, using the network topology and the credibility of users on their own characteristics so as to identify rumors even before they reach the stage of massive spread. These strategies emphasize the importance of combining content-based and propagation-based properties, as it is possible to assume that hybrid models are the most promising strategy in the context of the real world. Equally, when news articles have manipulated images or videos, multimodal detection methods that include visual, textual and video attributes are superior in detection accuracy. To illustrate a point, CNN-extracted visual characteristics with transformer-based textual ones can be used to achieve increased accuracy by as much as 7 percentage points over unimodal methods, which explains the rising significance of applying visual misinformation detection to textual documents as well.

Data analysis sheds light on the endemic constraints which hamper generalization. The most popular datasets, including LIAR, FakeNewsNet, PHEME, and BuzzFeed News, are strongly prejudiced in the content they include in terms of English language contents and Western socio-political backgrounds. This language and cultural bias raises important questions concerning the viability of the already established models to the global social media, where fake news can be presented in various languages, cultural backgrounds and in various stories. Further, the absence of multilingual or culturally homogenous data sets does not allow the creation of models that can sustain a high level of accuracy when used by the non-English speaking population.

The issue of ethical considerations and algorithm bias persist in the implementation of AI-guided fake news detection systems on a large scale. The review has found three main sources of bias: selection bias in training data, confirmation bias in human-labeled data and temporal bias with concept drift. These biases not only jeopardize the level of objectivity of the automated detection systems but also have a more far-reaching effect on society, such as possible stifled truth and unwilling bias of political or cultural nature. There is further a technical vulnerability of adversarial attacks in which a malicious agent can intentionally alter content to avoid detection, which implies that model robustness to adversarial attacks should be a focal point in model design.

In order to put these results into contexts, Table 1 provides the overview of the comparative performance and the specifics of various AI paradigms used in fake news detection:

AI Paradigm	Strengths	Limitations	Accuracy / F1-score range
Classical Machine Learning (SVM, Naive Bayes)	Simple, interpretable, fast on small datasets	Limited scalability, feature engineering required	72–81%
Deep Learning (LSTM, RNN)	Captures sequential dependencies, better context understanding	High computational cost, requires large datasets	82–87%
Transformer-based NLP (BERT, RoBERTa, XLNet)	Context-aware, high accuracy, handles large corpora	Computationally intensive, potential overfitting, sensitive to language bias	90–92%+
Graph-based / Propagation Models	Early detection, leverages network structure, user credibility	Requires propagation data, sensitive to network incompleteness	88–89%
Multimodal Detection (text+image/video)	Detects visual and textual misinformation, cross-modal consistency	Dataset and model complexity, computational cost	+3–7% improvement over unimodal

Table 1: Comparative summary of AI approaches for fake news detection

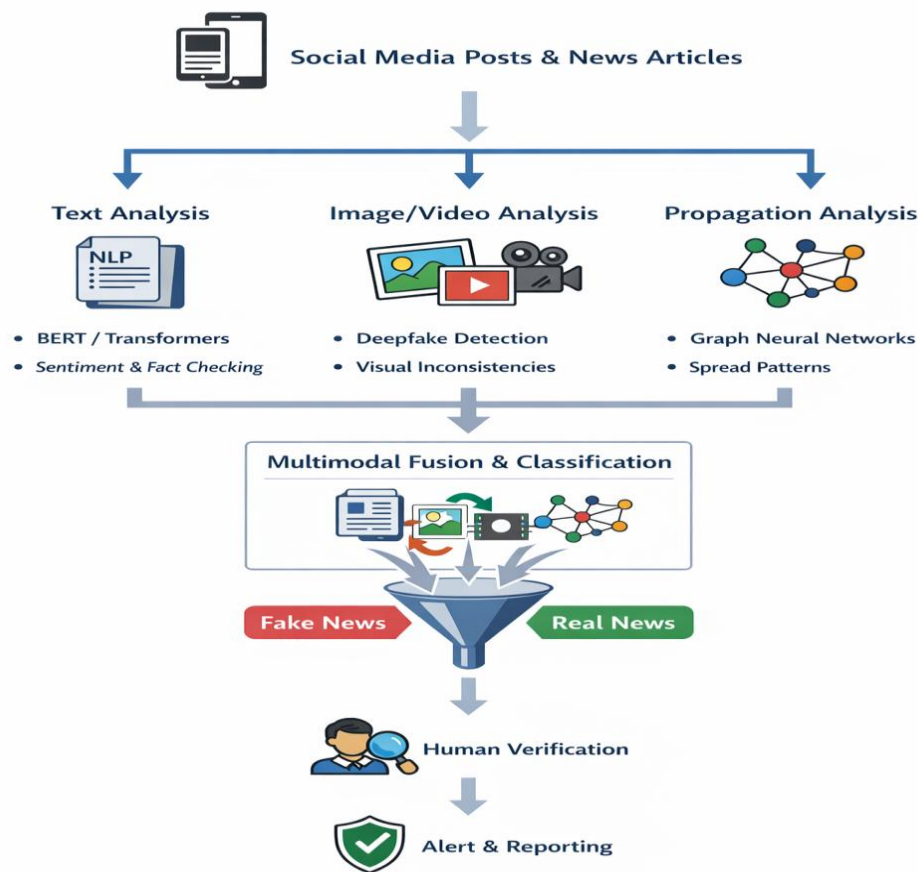


Figure 1: AI-Based Fake News Detection Pipeline

6. Conclusion and Policy Recommendations

6.1 Conclusion

This review has shown that artificial intelligence has significant potential as a means of identifying fake news on social media and transformer-based NLP models, graph neural networks, and multimodal fusion architectures have demonstrated a high level of performance on the available benchmark. The history of the field, starting with classical feature engineering, moving to pre-trained language models and graph-based propagation analysis, indicates the maturation of AI as a field, as well as its growing ability to deal with complex and socially relevant classification problems. Merging of NLP and computer vision together with the social network analysis in multimodal and hybrid detection systems is the new frontier in the area where the detection performance is much more robust compared to any single-modality system.

Nevertheless, significant gaps and limitations have also been noted in this review and have to be resolved before AI-based detection systems can be used responsibly on the scale. The overwhelming share of English-language data, the vulnerability of the existing models to adversarial manipulation, the issue of concept drift as fake news approaches evolve and under-investment in algorithmic bias and transparency are among the major obstacles to successful real-world implementation. The negative aspects of automated content moderation, especially the possibility of false positives leading to the suppression of legitimate speech, require immediate and long-term study on the part of researchers, developers of platforms, and policymakers. The introduction of AI as the judge of the truthfulness of information poses serious concerns regarding the question of accountability, democratic rule, and the acculturation of the knowledge base in the hands of a few technology firms (Wardle and Derakhshan, 2017).

The results of this review also emphasize the fact of the fundamentally interdisciplinary nature of the fake news problem. Even highly technical methods based on computing alone cannot be entirely used to explain the social, psychological, political, and cultural processes that fuel the production and consumption of misinformation. Subsequent studies should thus be pegged on actual cooperation between computer scientists, communication scholars, psychologists, political scientists, and ethicists. Multidisciplinary research groups are in a better position to come up with detection systems that are not only accurate but also sensitive to their context and to the cultures they affect, as well as to the democratic ideals.

6.2 Recommendations

Technical Recommendations

The main focus of the future research must be on the creation of multilingual and cross-cultural fake news datasets that are comprehensive and that reflect the linguistic and cultural diversity of the global social media. The recent focus on language detection with low resources, using cross-linguistic transfer learning and multilingual pre-trained models, including mBERT and XLM-R, is especially urgent due to the global nature of misinformation campaigns (Hardalov et al., 2022). Another priority that should be given by researchers is the creation of adversarially resistant detection models that can be resistant to attempts to deceive by the detection of deliberate attempts to confuse it, even by using synthetic text generated by large language models. The ability to update the architectures of continuous learning models to concept drift without losing the ability to detect future occurrences in a catastrophic way is a promising future direction in sustaining the performance of the detection models. Detection systems should be equipped with explainable AI (XAI) methods that can offer easily understandable explanations as to why the automated system has made a classification, which can allow humans to control automated moderation processes in a meaningful way (Pennycook and Rand, 2021).

Policy and Governance Recommendations.

The governance systems of platform should be revised to make AI-based content moderation systems contain mechanisms of independent algorithmic auditing, and the results thereof are publicly available. The datasets, methodologies, and performance metrics of their detectives should also be revealed to social media firms so that they would be subjected to outside review and responsibility. The types of regulatory frameworks that regulate the use of AI in content moderation like the Digital Services Act by the European Union should be expanded and reinforced to encompass the mandatory human review of disputed cases of content moderation and user-friendly appeals procedures in case of user-generated content being wrongly tagged. The governments also need to invest in the public media literacy strategies aimed at increasing the ability of citizens to assess the quality of information critically, which is why the technological methods of detecting misinformation should be supplemented with human interventions (Lazer et al., 2018). To create internationally relevant standards of AI-based misinformation detection that safeguard the freedom of expression and effectively address the disinformation harm, international coordination in the form of international bodies like UNESCO and the United Nations is required.

Funding: Funding is inappropriate for this research study.

Data availability: The statistics supporting the outcomes of this research are accessible upon reasonable request from the first author.

Declarations

Ethical approval: Not applicable

Consent to participate: Not applicable

Consent to publish: All authors have given consent to publish.

Competing interest: The authors have no competing interests to declare.

References

1. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236. <https://doi.org/10.1257/jep.31.2.211>
2. Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 549–556. <https://doi.org/10.1609/aaai.v34i01.5393>
3. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675–684. <https://doi.org/10.1145/1963405.1963500>
4. Cinelli, M., Quattrocioni, W., Galeazzi, A., Valensise, C. M., Brugnoti, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, 10(1), Article 16598. <https://doi.org/10.1038/s41598-020-73510-5>
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 1, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
6. Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020). The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys*, 53(4), 1–36. <https://doi.org/10.1145/3393880>
7. Hardalov, M., Arora, A., Nakov, P., & Augenstein, I. (2022). A survey on stance detection for mis- and disinformation identification. *Findings of the Association for Computational Linguistics: NAACL 2022*, 1259–1282. <https://doi.org/10.18653/v1/2022.findings-naacl.94>

8. Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. *The World Wide Web Conference*, 2915–2921. <https://doi.org/10.1145/3308558.3313552>
9. Kula, S., Choraś, M., & Kozik, R. (2021). Application of the BERT-based architecture in fake news detection. *Proceedings of the International Conference on Computational Science*, 239–249. https://doi.org/10.1007/978-3-030-77967-2_20
10. Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
11. Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions. *PLOS Medicine*, 6(7), Article e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
12. Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6412–6421.
13. Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
14. Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
15. Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 22–32. <https://doi.org/10.18653/v1/D18-1003>
16. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 1–11. <https://doi.org/10.1109/ICCV.2019.00009>
17. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797–806. <https://doi.org/10.1145/3132847.3132877>
18. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM transactions on intelligent systems and technology (TIST)*, 10(3), 1-42. <https://doi.org/10.1145/3305260>
19. Shi, B., & Weninger, T. (2016). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104, 123–133. <https://doi.org/10.1016/j.knosys.2016.04.015>
20. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. <https://doi.org/10.1145/3137597.3137600>
21. Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312–320. <https://doi.org/10.1145/3289600.3290994>
22. Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153. <https://doi.org/10.1080/21670811.2017.1360143>
23. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. *Proceedings of the 2018 Conference of the North American*

- Chapter of the Association for Computational Linguistics, 1, 809–819. <https://doi.org/10.18653/v1/N18-1074>
24. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>
25. 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>
26. Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2, 422–426. <https://doi.org/10.18653/v1/P17-2067>
27. Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*, Strasbourg: Council of Europe, 27, 1-107.
28. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in Neural Information Processing Systems*, 32, 9054–9065. <https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html>
29. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40. <https://doi.org/10.1145/3395046>