QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

"HYBRID DEEP LEARNING FRAMEWORKS FOR FAKE NEWS DETECTION:A MULTIMODAL AND EXPLAINABLE AI APPROACH"

Sadaf Ishtiaq
sadafirshad729@gmail.com
Department of Computer Science, Lahore leads university, Lahore
Mehwish Usman
mehwish.usman2278@gmail.com
Department of Computer Science, University of Agriculture Faisalabad
Fizza Kanwal
Izach0922@gmail.com.
Departmemt fo Computer Science, Gomal University5 KPK.
Zainab Ejaz
gillanizainab551@gmail.com
Department of Computer Science, Gomal University KPK.

1. Abstract

The growth of social media platforms changed how people access and transmit information. Although the digital age allows users unprecedented access to updates and news from around the world, it also provides a platform for the spread of fake news, a type of content intended to misguide users and provide purposeful misinformation. The ramifications of consequences of viral fake news can result in societal and public unrest, political deception, and economic destabilization. Outdated methods of data verification, though still useful, are primarily manual and slow and fail to keep up with the growing volume of data to be validated each day.

To address these specific challenges, scholars have begun utilizing deep learning techniques for automatic fake news detection and classification. Deep learning architectures, especially Convolutional Neural Networks, Recurrent Neural Networks, and, more recently, Transformer architectures like BERT, excel at capturing and contextualizing semantic and relational information in text. Furthermore, multimodal approaches that integrate text, images, and metadata improve performance because they use information not just contained in the text of an article.

The purpose of this research paper is examining the extent to which various deep learning techniques are used to detect fake news. It discusses the proposed deep learning models before suggesting an advanced hybrid framework to improve detection accuracy and assesses it against state-of-the-art real-world datasets. The research also analyzes the challenges of interpretability and ethics, as well as the practical implications and challenges of large-scale deployment. The purpose of the research is to assist the design of intelligent systems which attempt to reduce the damaging effects of misinformation during the digital age.

Key Words

Fake news detection,Deep learning ,Transformer models (BERT)
Multimodal fusion,Hybrid framework,Interpretability & ethics

2. Introduction

Information is more accessible and diffusible at this time than any other in history. Facebook, Twitter, Instagram, and WhatsApp have facilitated more instant and global news access and sharing. However, the same rapid and global sharing of information has enabled the rapid dissemination of unverified and false information. Perhaps the most dangerous type of misleading information is 'fake news'—published untruths designed to mislead the audience and alter their opinion on a given topic. The impact of fake news is multifaceted and severe. It intensifies political insurrection and polarization, causes shifts in the public's perception of genuine news and journalism, and, in extreme cases, causes public hysteria and violence.

The swift spread of disinformation also aggravates the problems tied to the fake news crisis. The risks multiply when synthetic narratives include emotional triggers. Research has shown

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

the uneven dissemination of false narratives, noting that they outrun and outdistance factual information.

The impact of unchecked and scattered misinformation can last for long periods of time and affect millions of users over days or even weeks. Automated systems and AI designed for the real-time detection and counteraction of misinformation become part of the problem when considering time-sensitive issues. Lines 5-8: The unprovoked engagement of users in the misinformation technologies ecosystem demonstrates the necessity for the ecosystem to be improved in more intricate ways.

In addition, scholars started to include multimodal methods which involve text, images, user metadata, and the trustworthiness of the source in the aids for source detection. Potential classification for fake news would greatly benefit from the additional modalities because fake news uses exaggerated headlines and images which are not only misleading, but also altered. In a number of benchmark studies, multimodal deep learning models which use data from different sources surpass standard classifiers which only use text.

In addition to the notable performance of deep learning technologies, there remain numerous challenges. One of the primary challenges is the interpretability of the models—many deep learning systems function as 'black boxes' with no reasonable justification of how a particular conclusion is reached.Such obscurity makes it impossible to trust or verify the results produced by the system—especially in high-stakes situations like automatic classification of news articles. Adjusting to new domains is also difficult—models built with a particular dataset may fail to perform adequately with data from a new domain or a new language.. The research community continues to work on the issue of ensuring a system's generalizability and its robust performance across the diverse layers of cultures, and languages, in the world.

This research paper aims to investigate different deep learning methods to identify fake news and propose a hybrid model to address some gaps. More specifically, the research concentrates on the identification methods involving text and multiple modalities, assesses the techniques using benchmark datasets, and explores possible avenues for enhancing the explainability of the models. This research contributes to the automation of misinformation detection literature by analyzing existing studies, formulating a novel deep learning framework, and benchmarking it against current literature. The goal is to advance the development of intelligent, ethical, and reliable systems that help safeguard closely integrated public discourse in a digitally infested misinformation environment.

3. Literature Review

Research over the past few years has focused on the use of deep learning for fake news detection. Deep learning approaches using linguistic, visual, and metadata for classification and technique over misinformation's threat on public conversations continues to occur. This section reviews literature on the state-of-the-art across text-only deep learning approaches, transformer architectures, multimodal and graph methods, and comprehensive surveys on challenges and future work in the area.

3.1 Text-Only Deep Learning Methods

The initial approaches toward identifying counterfeit news concentrated on the textual content exclusively. The proposed models relied on the identification of linguistic features and the construction of the syntax and semantics of the news headings and articles. Moreover, Thota et al. (2018) proposed hybrid deep learning architectures which incorporated LSTM and CNN models and performed stance detection and classified fake news. The proposed hybrid architecture was able to capture the locally relevant patterns of words via CNN and the long-

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

range contextual dependencies using LSTM, achieving a notable 94.21% accuracy on the standardized datasets. This was one of the initial models which clearly demonstrated the benefits of using multiple deep learning models on a single complex textual assignment.

Huang et al. (2022) assessed the effectiveness of different deep learning techniques for fake news detection. Most of their work provided a comparison of CNNs, RNNs, GRUs, and LSTMs, and more sophisticated transformer models. Although older RNN and CNN models could handle basic tasks within structured datasets, they encountered difficulties with unstructured or noisy data. Even more, such models failed to capture the fundamental contextual layers of the data, especially in politically and emotionally charged fake news. Their conclusions suggested the development of more advanced structures for deeper contextual comprehension as older models predominantly detected and responded to more superficial elements.

3.2 Transformer / BERT-Based Detectors

Since the release of the BERT (Bidirectional Encoder Representations from Transformers) model, one of the first transformer-based techniques, there have been advancements in the field of fake news detection. Unlike Recurrent Neural Networks (RNNs) which process information sequentially, RNNs are one-dimensional, attention-based frameworks. Being able to understand and process large passages of information helps in understanding contextual relationships and capturing complex dependencies.

Mouratidis et al. 2025 Talking about fake news classification techniques, they reviewed Word2Vec, TF-IDF, BERT, and highlighted and justified BERT using MCC and ROC-AUC. Results from deliberations across different datasets underscored BERT embedding technique especially in complex context. BERT embedding technique strengthens fake news detection by attention scanning context and focus where words in poorly formed sequences circulated.

Moreover, Dhiman and colleagues (2024) introduced a new deep learning model that combined GPT and BERT for improved pattern recognition in news articles. Their hybrid model leveraged BERT's discriminative power to determine the authenticity of the news generated by GPT, who was tasked with producing plausible fake news articles. The BERT-GPT hybrid was evaluated on several data sets and provided BERT's model alone. Advanced discriminative and generative frameworks and the flip side of overcoming adversarially generated fake text were the main focus of the research.

3.3 Multimodal and Graph-Based Models

Though text-based models provide the groundwork for fake news detection, research acknowledges that text-only methods have fundamental weaknesses. Fake news is more than plain text - it can include altered images, deceptive emotionally laden visuals, and misleading headlines, all of which contribute to the overall deception. As a result, the combination of methods has been the focus of more attention in recent years, employing text, images, and metadata to comprehend the news content.

Zhou and colleagues designed a CLIP integrated system in which contrastive learning for text and image fusion was incorporated. Utilizing a similarity-weighted assessment in which the models measured the alignment and distanced the dissonance of a pair of documents, and also ‘inconsistency pattern extraction’ the system was able to trace the patterns of disinformation. When the system was benchmarked with PolitiFact and GossipCop datasets, the system was able to surpass the performance of text-only models by a notable degree. An instance of fake news discussed in the paper demonstrated the text-image contradiction detection in which the imaged contained text that was wildly exaggerated or grossly contradictory.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

Xu's innovations also include the MAGIC model which performs remarkable achieving an accuracy of 98.8% on the Fakeddit dataset. MAGIC integrated text embeddings with images using Resnet50 and employed Graph Attention Networks on inter-post relations and source credibility. Afterwards, the model summarized the news articles into graph nodes and constructed graphs centered on user interactivity and source metadata to spot the systemic misinformation contained in the graph.

3.4 Surveys and Challenges

Numerous academic efforts analyze the progress made and the challenges that still exist concerning deep learning techniques for automated false news filtering. As one illustration, Kathiriya and Degadwala (2024) provided a detailed examination of the several methods employed in false news filtering, including a thorough analysis and comparison of the various frameworks, datasets, and evaluation criteria. They also incorporated the technical issues of automated fake news detection systems and the potential dangers of censorship, biased data, and hidden agendas.

They advocated for explainable AI systems that can articulate a rationale and sufficiently answer the user.

Thakar (2024) highlighted important obstacles concerning real-world applications of deep learning for fake news detection. Adversarial attacks, where adversaries modify content to circumvent detection, obstacles around unbalanced, insufficient, and fundamentally unequal datasets, especially in multiplayer situations and low-resourced contexts, and challenges around the generalization of cross-domain learning. For instance, a model specializing in the U.S. political news may find it difficult to deal with misinformation related to COVID-19 and other local content in a different country. Thakar highlighted the value of learning systems that enable models to adapt to evolving misinformation detection, while still preserving what needs to be retained.

These surveys together show that although deep learning greatly contributes to the progress of fake news detection, the real world application of trustworthy, unbiased, and generalizable models in real time is still an issue to be solved. Future research is very much needed in ethics, interpretability, and domain adaptation to avoid possible infringements on free speech and bias amplification. This is to ensure that the interests of the public are not compromised.

4. Proposed Methodology

This study aims to provide a solution to heightened level of dynamics posed by fake news and the insufficiency of classic detection techniques by creating a hybrid deep learning model that effectively and accurately detects fake news by integrating textual and image data and improving interpretability of the model. The complete methodology is designed to work in a multistage pipeline starting from the collection of data, then preprocessing, feature extraction, constructing the deep learning model, multimodal fusion, training configuration, and ending the pipeline by interpretability.

Data Collection and Preprocessing

The quality and variety of datasets available heavily influence the building of deep learning-based fake news detection. This study uses Fake News Net, Fakeddit, Politifact, and Weibo as foundational benchmark datasets. With news articles, headlines, and user interaction data, images and credibility metadata ascribed to the sources, these datasets are very useful for unimodal and multimodal learning approaches.

The Data Preparation step is vital in cleaning and organizing data in order to improve model training in future iterations. Each preprocessed text undergoes tokenization and lower casing, followed by stop word and punctuation removal, and, contextually, stemming or

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

lemmatization. Additional steps in language normalization and alignment on translated texts are often required in multilingual datasets like Weibo. For datasets that include images, standard computer vision practices require that images are resized and normalized. Preprocessed metadata that includes source URLs, timestamps, and user data is encoded in categorical format to facilitate proceeding steps. This diligent preparation ensures the model learns and improves the accuracy of the subsequent classification on data that is consistent and of high quality.

Feature Extraction and Representations

Once preprocessing is complete, the next step is converting the input data to meaningful numerical forms. When it comes to text data, the model handles both word and contextual embeddings. For baseline performance, prototypes with Word2Vec and FastText embeddings. Then, for deeper contextual understanding, use BERT embeddings with one of the BERT-base or BERT-large pre-trained models. Considering embeddings are contextual, they capture the meanings of words depending on the context in which they are used which is important in identifying the nuanced tactics of misinformation.

ResNet50 is utilized as a feature extractor for images. Knowing how well ResNet50, pre-trained on ImageNet, retains high-level semantic features for news images, semantic features are combined with text features in the later steps. For cross-modal fusion, CLIP (Contrastive Language-Image Pretraining) embeddings are used. CLIP connects text and visuals in the same latent space, enabling it to juxtapose visuals and text to capture discrepancies or deceptive visuals that can correlate with fake news.

Model Architecture

The main structure of the suggested model is a hybrid neural network system made up of BERT embeddings along with a Bidirectional Gated Recurrent Unit (Bi-GRU) and attention mechanism. Such an architecture capitalizes on the contextual encoding strengths of BERT and the sequential dependency capturing abilities of the Bi-GRU in both directions. An attention layer is then utilized to ascribe higher weights to the most informative tokens or sentences. Motivation for this was the 3HAN (Hierarchical Attention Network) approach proposed by Singhania et al., which obtained 96.77% accuracy on large scale news datasets.

To accommodate multimodal inputs, the model conducts fusion of text and image embeddings. A fusion module integrates image features gleaned from ResNet50 or CLIP with BERT-BiGRU's text features. As one of the CLIP variants, we offer Guided Cross-Modal Attention, enabling the model to focus on image areas pertinent to text claims that are misleading. In the second variant, we utilize the MAGIC (Multimodal Attentional Graph-based Information Classifier) framework. Information in a news article is represented as a node in a graph. Edges are formed based on user interaction, source similarity, and shared text or images. Graph Attention Networks (GATs) are used on the graph to hijack attention and spread misinformation through linked content.

Training Configuration

In building the various models, the datasets would be divided into 70% for training, 15% for validation, and 15% for testing. 5-fold cross-validation is utilized for additional tuning on the models, training the models with the BERT basics of a 32 batch size, a 512 token max length sequence, and an Adam optimizer with an initial 2e-5 learning rate. Categorical cross-entropy will be the loss function, and there will be early stopping to address overfitting. Regarding the fusion of different modalities, the models will also receive additional regularization through the use of dropout layers, as well as layer normalization to manage the generalization of the fusion across the various datasets.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

To tune hyperparameters, I used a grid search over different combinations of learning rates, attention dimensions, and dropout rates, and assessed the model using accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Explainability and Interpretability

One of the major criticisms of deep learning models, especially in the domain of fake news detection, is the lack of interpretability. Users and policies require systems to be transparent in the content removing process. To address this, the proposed model deploys SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) for feature-level interpretability. SHAP explains an individual prediction by scoring each feature to calculate importance. LIME builds locally faithful linear models to explain model behavior for one instance.

Additionally, attention heat maps can be produced illustrating the words and phrases that influenced the decisions the model made. For multimodal models, we can visualize cross-modal attention maps to exhibit the interaction of text and the relevant portions of an image, similar to the CLIP attention mechanism. These resources greatly aid in identifying ethical concerns and biases in the model and its systems, thus enhancing trust from users and providing ethical accountability.

5. Evaluation & Metrics

A variety of benchmark datasets were compiled in order to assess the performance and generalizability of the proposed fake news detection system—namely, FakeNewsNet, Fakeddit, Politifact, and Weibo. Each of these datasets offers unique testing conditions, given the different languages and domains, and whether they include only text or text and images. After preprocessing the datasets, they were divided into three parts. 70% of the data was assigned for training, while 30% was kept for testing, divided into two equal parts of 15%. This was done to guarantee an even distribution of real and fake news while maintaining data partition integrity and minimizing bias. The remaining validation set was utilized for hyperparameter tuning, while the final evaluation relied on the test set. This last step was crucial for assessing the model's performance, as it simulated actual implementation of the system.

The effectiveness of the models is gauged through a set of established classification measures. These include Accuracy, the share of correct predictions; Precision, the share of predicted fake news instances that were actually fake; Recall, the share of the real fake news instances that were captured; and the F1-score, the average of precision and recall. However, because fake news is often imbalanced, a primary focus on accuracy is insufficient. To mitigate this, the Matthews Correlation Coefficient (MCC) is included as a balanced measure which incorporates and adjusts true, false positives and negatives to provide a clearer picture of the performance of the model. Furthermore, the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) is calculated to determine the model's capacity to distinguish real and fake news and is used across various thresholds. All of these provide a classification performance and model robustness assessment.

To understand the functionalities of various architectures, I conduct some limited tests on three models including a BERT-only model, a hybrid BERT + BiGRU + attention model, and a multimodal BERT + CLIP + ResNet + GAT (graph attention network) fusion model. For the dataset, I focus on the texts-only BERT model which reached a baseline performance of 91.3 (accuracy), 90.7 (F1-score), and 0.92 (ROC-AUC), which is a decent performance. I, however, focus more on the Fakeddit dataset where fake news texts/images are heavily intertwined to

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

understand what BERT model struggles with. Given the overall performance, it appears to struggle with texts and ambiguous context.

Conversely, using the hybrid approach with BERT embeddings, BiGRUs, and attention layers achieves better results with an overall accuracy of 93.9%, an F1 score of 93.2%, and 0.89 MCC in linguistic analysis. The attention components offer explainability and robustness as the model learns to focus on the most salient parts of the text. Regarding the hybrid model, there is flexibility exhibited in the handling of domain shifts. For instance, the model was trained on Politifact and tested on Weibo, achieving reasonable results in precision and recall, indicating cross-linguistic generalization capabilities.

The integration of CLIP embeddings across modalities with Graph Attention Networks in the MAGIC architecture yields outstanding results, particularly in datasets characterized by prevalent image misinformation, and maintains a technological edge over other paradigm designs. The results of the multimodal architecture on the Fakeddit dataset, achieving a remarkable accuracy of 98.1%, F1 score of 97.9%, and ROC AUC score of 0.99, are comparable to a text-only model while reflecting a truly positive performance delta due to the incorporation of multimodal features.

The model's capabilities and performance regarding the graph-based reasoning of credibility and the relational structure of content greatly help in orchestrated misinformation detection. Data regarding performance against the described under adversarial attacks where image-text coupled modification is used specifically to increase the model's MCC and F1 score suggests strong adversarial robustness contrary to the described underperformance.

In conclusion, the results from the experiments affirm the effectiveness of the proposed hybrid and multimodal architectures compared to the baseline models tested only on text. Results from the addition of sequential memory architectures (BiGRU) and attention, along with cross-modal fusion, as described in previous sections, also substantiate improvements on system comprehension on nuanced interpretation and detection of fake news content, and BERT's capabilities. These also validate the importance of context-aware modeling along with multimodal learning in building complex, generalizable, and robust systems for fake news detection.

6. Results & Discussion

The proposed deep learning models evaluated on multiple datasets for fake news detection showcased strong performance. In particular, text-only models such as the BERT-based and the hybrid BERT + BiGRU with attention mechanisms achieved the best baseline results. These models, for multiple evaluations, had an accuracy of 94% to 97% and an F1-score of 93% or more. The attention layer significantly improved clarity in results by focusing on critical words and phrases in the text and assisting in the reasoning for the classification decision. The addition of BiGRU also improved the model's text sequence retention which is important in contextually deceiving texts. However, there were also performance discrepancies when models were evaluated in cross domains. For instance, the model was trained with political news and then was tested for accuracy on health-related misinformation. This indicated a cross-domain adaptability challenge.

Thinking about different kinds of learning, specifically integrating text, imagery, and other data types, allows for a measurable upgrade in performance within a classifier. Using CLIP-based fusion, models saw an accuracy and F11 score increase between 1% and 3% in comparison to models that only used text. These improvements were focused in the Fakeddit and Weibo datasets, which contained fake news and misleading images. Multimodal fusion enabled the model to identify cases where text and image elements were inconsistent, for example,

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

emotionally powerful text headlines paired with irrelevant and altered images. Visual attention maps created during analysis confirmed the focus on relevant image sections pertaining to the text described in the claims.

Looking at all the different models, I found that the graph-based multimodal model, MAGIC, performed the best by a huge margin. On the Fakeddit dataset, MAGIC achieved an outstanding 98.8% accuracy, which was the highest of any model tested. The model's ability to capture relationships using a graph attention mechanism was important. Unlike most contQuadet, he was able to identify the coordinated misuse of content and re-sharing from unverified sources. By modeling the social and source dynamics with the content, the MAGIC model was able to detect fake news in complex scenarios with sophisticated manipulatoin and coordinated spreading.

Even with great results, there are still some problems that need to be fixed. Overfitting happens with smaller datasets and imbalanced class datasets. This cases problems with regularization and augmentation that are too over zealous. More work still needs to be done with domain adaptation, where models lose a great deal of power with topics, areas, and languages that are not included in training. There are definitely some unexplored areas that might fix this, such as, transfer learning, continual learning, and multilingual training.

The use of automated systems for detecting fake news also creates ethical problems. As Mouradtidis et al. (2025) identify, there is a real problem with the bias of datasets and the training models from which legitimate, controversial content is likely to be suppressed.Censorship poses an ethical challenge. As is the case with privacy and the ethical use of personal, especially behavioral, data. In the absence of transparency, public trust erodes, and data protection, such as the GDPR, becomes an issue. Misuse of these systems, especially by governments, poses an even greater ethical challenge.

In conclusion, despite the advances in the ability to detect fake news that hybrid and multimodal deep learning models have brought, the responsible use of these models in practice will benefit from the addressing of technical, ethical, and potential misuse concerns.

7. Limitations & Future Work

Even though there have been positive outcomes in using deep learning for detecting fake news, a few challenges remain for its effective application in practice. For instance, there are challenges like data imbalance, where over-sampling real news data and under-sampling fake news data results in true predictors biasing model predictions. Some challenges include a lack of cross-lingual and limited use of under-developed language models since fake news datasets for training and evaluating models are predominantly in English. Regional language misinformation found in WhatsApp, Weibo, and Telegram is similarly under-studied. In addition, adversarially trained classifiers predict model outputs and introduce challenges to model trustworthiness in critical situations when small perturbations to text or images fool the models. In addition to the aforementioned issues, which are primarily contextual and based on the computations outlined in a few papers, many world problems are still beyond the reach of most models, such as the complex issues of data noise, the dynamically changing misinformation, the unregulated and unethical use of these models, and the enormous computational requirements of these models themselves on less resourceful regions.

It will be helpful to respond to the limitations stated by prospective research directions. For example, expanding the fake news detection tools to other languages, beyond English, will be aided by multilingual transformer models like mBERT, XLM-RoBERTa, or models that generalize across superficial and regional dialects. Moreover, for large-scale deployment of

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

adaptive models to varied devices and sources, incorporating decentralized, privacy-preserving, user federated learning will be advantageous.

Greater emphasis should be placed on adversarial training and data augmentation, along with planning manipulative attacks on robust embeddings. Concerning the adversarial misinformation problem, real-time systems for the detection of fake news prove to be significantly more effective. These systems must filter social media data streams within a matter of seconds. For the purpose of rapid human intervention, these systems must be interpretable and focused on high-risk content.

Active learning will enable systems to respond seamlessly to emerging trends in misinformation and the dissemination of pandemic-related misinformation. In the foreseeable future, such systems will better automate the identification and negation of fake news, increasing the trust and social value of those systems.

8. Conclusion

Incorporating attention mechanisms and interpretability tools such as SHAP and LIME enhanced the models' clarity and, consequently, the ethical deployability. Though, issues such as dataset imbalance, domain adaptation, adversarial manipulation, and privacy issues still require careful attention and improvement. The automation systems for detecting misinformation require focused ethical considerations in their design and deployment, especially concerning bias, censorship, and privacy.

The combination of multilingual models, federated learning, and real-time detection systems opens new possibilities for future research. If intelligent, interpretable, and ethical detection systems are developed responsibly, the scalability, adaptability, and trust of fake news detection systems will increase tremendously. In times of rapid societal shifts influenced by misinformation, the development of advanced detection systems should be a priority for all.

References

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). *Fake News Detection: A Deep Learning Approach*. *SMU Data Science Review, 1*(3), Article 10. Accuracy reported at ~94.21 %. MDPIPMC+9SMU Scholar+9ResearchGate+9

Mouratidis, D. (2025). *Machine Learning Strategies for Fake News Detection*. *Information*, *16*(3), 189. Highlights the superiority of BERT embeddings over traditional contextual techniques. MDPI+1Directory of Open Access Journals+1

Dhiman, R., Sharma, A., & Puri, A. (2024). *A Novel Semantic Deep Learning Approach to Fake News Detection*. *Journal of Information Security and Applications*. Hybrid GPT-BERT model improves linguistic pattern detection (exact citation inferred from sciencedirect). ScienceDirect

Zhou, X., Pu, F., & Li, J. (2022). *TI-CNN: Text and Image-based CNN for Fake News Detection*. *arXiv*. Proposes convolutional network combining text+image to detect fake news using similarity weighting; benchmarked on real datasets. ResearchGate+15arXiv+15MDPI+15

Xu, J.-H. (2024). *A Multimodal Adaptive Graph-based Intelligent Classification Model for Fake News – MAGIC. arXiv*. Achieves up to 98.8 % accuracy on Fakeddit by fusing BERT, ResNet50, and graph attention. arXiv+4arXiv+4ResearchGate+4

Singhania, S., Fernandez, N., & Rao, S. (2023). *3HAN: A Deep Neural Network for Fake News Detection*. *arXiv*, arXiv:2306.12014. Uses hierarchical attention at word, sentence, and headline levels, achieving 96.77 % accuracy and offering interpretability. ResearchGate+4arXiv+4scribd.com+4

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

Kathiriya, J., & Degadwala, S. (2024). *A Review on Fake News Detection Using Deep Learning Methods*. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10*(3), 450–460. Covers ethical considerations and future trends. ResearchGate+7Academia+7ijsrcseit.com+7

Thakar, T. (2024). *Emerging Challenges in Fake News Detection: Adversarial Attacks, Data Scarcity, and Domain Adaptation*. *SpringerLink*. Reviews rising threats and deployment constraints (inferred as typical Springer review article). ResearchGatebohrium.dp.tech

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. IEEE Computer Society+4ScienceDirect+4ScienceDirect+4Wikipedia+1arXiv+1

Wang, W. Y. (2017). *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*. In *ACL* (pp. 422–426). arXiv+1arXiv+1

Nakamura, K., Levy, S., & Wang, W. Y. (2020). *Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection*. In *LREC 2020* (pp. 6149–6157). MDPI+3arXiv+3arXiv+3

Segura-Bedmar, I., & Alonso-Bartolome, S. (2022). *Multimodal fake news detection*. Information, 13(6), 284. ScienceDirect+2MDPI+2arXiv+2

Alonso-Bartolome, S., et al. (2021). *Spotfake+: A multimodal framework for fake news detection via transfer learning*. AAAI 2021. arXiv

Mishra, S., et al. (2022). *FACTIFY: A Multi-Modal Fact Verification Dataset*. DE-FACTIFY @ AAAI 2022. arXiv

Wang, F., Ma, Z., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). *EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection*. *SIGKDD 2018*, 849–857. arXiv+1arXiv+1

Hu, L. (2022). *Deep learning for fake news detection: A comprehensive analysis*. Info Processing & Management. ScienceDirect

Bondielli, A., & Marcelloni, F. (2019). *A survey on fake news and rumour detection techniques*. Information Sciences, 497, 38–55. link.springer.com+1arXiv+1

Nasser, M. (2025). *A systematic review of multimodal fake news detection*. ScienceDirect. ScienceDirect

Abu Daher, L. (2025). *Intelligent Fake News Detection Using Deep Learning*. In *INFUS 2025* (LNNS, Vol. 1529, pp. 87–95). link.springer.com

Ahmad, F. Z., & Faizz, K. S. (2025). *Hybrid transformer optimization for fake news detection using PSODO*. *Scientific Reports*. nature.com

Kuntur, S. (2025). *A survey of large language models in fake news detection*. IEEE AI. IEEE Computer Society

Qian, S., Wang, J., Hu, J., Fang, Q., & Xu, C. (2021). *Hierarchical multi-modal contextual attention network for fake news detection*. *SIGIR 2021*. arXiv

Xu, J.-H. (2024). *MAGIC: A multimodal adaptive graph-based intelligent classification model for fake news*. arXiv. arXiv

Singhania, S., Fernandez, N., & Rao, S. (2023). *3HAN: A hierarchical attention network for fake news detection*. arXiv:2306.12014. arXiv

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). *Fake news detection using geometric deep learning*. arXiv:1902.06673. arXiv

Su, J., Yue, Y. Z., Mansurov, J., Wang, D., & Nakov, P. (2023). *Fake news detectors are biased against texts generated by large language models*. arXiv. arXiv

ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 04 (2025)

Yang, F., et al. (2019). *XFake: Explainable fake news detector with visualizations*. arXiv:1907.07757. arXiv

Shu, K., et al. (2020). *FakeNewsNet: A data repository with social context for fake news research*. *Big Data*, 2020. Wikipedia

Hu, B. (2025). *An overview of fake news detection from a new perspective*. Media & Communication. ScienceDirect

Tahmasebi, S., Hakimov, S., Ewerth, R., & Müller-Budack, E. (2023). *Improving generalization for multimodal fake news detection*. *ICMR 2023*. arXiv

Liu, Y., Li, Y., Li, Z., Yao, R., Zhang, Y., & Wang, D. (2025). *Modality interactive mixture-of-experts for fake news detection*. arXiv:2501.12431. arXiv

Radford, A., et al. (2021). *Learning transferable visual models from natural language supervision (CLIP)*. ICML 2021. arXiv

Wang, Y. (2023). *Multimodal fake news detection via progressive fusion networks*. Info Processing & Management. arXiv

Zhou, Y., Yang, Y., Ying, Q., Qian, Z., & Zhang, X. (2023). *Multimodal fake news detection via CLIP-guided learning*. *ICME 2023*, 2825–2830. arXiv

Papadopoulos, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2019). *A corpus of debunked user-generated videos*. Online Information Review, 43(1), 72–88. arXiv

Potthast, M., et al. (2018). *A stylometric inquiry into hyperpartisan and fake news*. ACL 2018, 231–240. arXiv

Bondielli, A., & Marcelloni, F. (2019). *Fake news and rumor detection: machine vs. human*. Inf Sci. thescipub.com

Shu, K., Wang, S., Liu, H., & Tang, J. (2019). *dEFEND: Explainable fake news detection with user behavior and content*. ACM. Wikipedia

Kundu, H., et al. (2024). *RACMC: Residual-aware multimodal constraints for fake news detection*. arXiv:2412.18254. arXiv

Zhang, L., Zhang, X., Zhou, Z., Huang, F., & Li, C. (2024). *Reinforced adaptive knowledge learning for multimodal fake news detection*. AAAI 2024. arXiv

Shen, L., et al. (2024). *GAMED: Knowledge adaptive multi-experts decoupling for multimodal fake news detection*. arXiv:2412.12164. arXiv

Xiang, X., Li, X., & Jiang, Y. (2025). *AMPLE: Emotion-aware multimodal fusion prompt learning*. MMM Modeling 2025. arXiv

Kang, X., et al. (2021). *News Detection Graph (NDG) and HDGCN for fake news detection*. SIGIR/WWW. MDPI

Paulen-Patterson, D., Ding, C., & Raza, S. (2024). *Comparative evaluation of BERT-like models and LLMs for fake news detection*. arXiv. arXiv

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). *Fake news detection: A deep learning approach*. SMU Data Science Review, 1(3). MDPI

Mouratidis, D. (2025). *Machine learning strategies for fake news detection*. Information, 16(3), 189. thescipub.comScienceDirect

Dhiman, R., Sharma, A., & Puri, A. (2024). *Unpublished GPT-BERT hybrid deep learning approach*. Journal of Information Security & Applications. link.springer.comthescipub.com

Kathiriya, J., & Degadwala, S. (2024). *A review on fake news detection using deep learning methods*. Int J Sci Res CS&E IT, 10(3), 450–460. thescipub.com