



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XII Month of publication: December 2025

DOI: <https://doi.org/10.22214/ijraset.2025.76628>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

A Review on Fake News Detection and Personalized Recommendation on Social Media using BERT

Adithya AA¹, Hana Habeeb Rahman², Goutham Krishnan³, Sylasree VA⁴, Asst.Prof. Reshma P D⁵
Department of Computer Science and Engineering Universal Engineering College Thrissur, Kerala, India

Abstract: *The rapid spread of misinformation on social media platforms poses a serious threat to public trust, social stability, and digital well-being, making fake news detection a critical research problem. To address this challenge, the integration of intelligent detection models with privacy-preserving mechanisms has become increasingly important. This paper presents a comprehensive review of transformer-based fake news detection approaches, particularly those utilizing Bidirectional Encoder Representations from Transformers (BERT), along with Federated Learning techniques for secure and decentralized personalization. The study analyzes state-of-the-art methods, benchmark datasets, and key performance metrics, demonstrating how BERT-based models effectively capture deep contextual semantics to accurately distinguish fake content from factual information. In parallel, federated learning enables distributed model training while preserving user privacy, making it suitable for deployment in recommendation systems on social media platforms. Furthermore, this review discusses a hybrid architectural perspective that combines a centralized BERT-based classifier with a lightweight, device-side federated recommender to deliver trustworthy and personalized news feeds. Finally, major challenges such as data imbalance, model explainability, and adversarial manipulation are examined, and future research directions are outlined toward robust, interpretable, and ethically aligned artificial intelligence systems for combating misinformation.*

Index Terms: Fake News Detection, BERT, Federated Learning, Privacy, Recommendation Systems, Social Media

I. INTRODUCTION

Social media platforms have fundamentally transformed the way people consume and share information. News, opinions, and trending stories can spread across networks within minutes, reaching millions of users with unprecedented speed. While this rapid dissemination enhances connectivity and awareness, it has also created fertile ground for misinformation and fake news. Inaccurate or intentionally deceptive content can manipulate public opinion, distort social discourse, and erode trust in legitimate information sources [5], [13], [14]. Traditional moderation strategies and manual fact-checking mechanisms are often too slow and resource-intensive to cope with the scale and velocity of modern social media streams [9], [16].

To address these challenges, researchers have increasingly turned to advanced machine learning techniques for automated fake news detection. Early approaches relied on handcrafted lexical, syntactic, and semantic features processed by classical machine learning algorithms such as Support Vector Machines, Decision Trees, and Random Forests [5], [13], [14]. Although these methods demonstrated reasonable performance under controlled conditions, they struggled to generalize across domains and adapt to evolving misinformation patterns.

The introduction of deep learning, particularly transformer-based architectures, has significantly enhanced detection capabilities. Bidirectional Encoder Representations from Transformers (BERT) and its variants provide rich contextual embeddings that capture subtle linguistic cues and long-range semantic dependencies [1], [6], [7]. Ayyub et al. [6] proposed a blended BERT-based approach that combines contextual embeddings with task-specific fine-tuning to improve classification stability, while Farokhian and Rafe [7] employed dual BERT networks to capture complementary contextual signals. Comparative studies by Chiang et al. [1] further emphasize BERT's superiority over traditional neural network models in source credibility recognition.

Beyond text-based models, ensemble learning and graph-based techniques have been introduced to enhance robustness by incorporating social context. Al-shaqi et al. [2] demonstrated that ensembles of diverse classifiers improve performance across heterogeneous datasets, particularly under noisy or adversarial conditions. Han et al. [3] applied Graph Neural Networks (GNNs) with continual learning to exploit relational structures in social media, highlighting the importance of user-post interactions and propagation patterns for accurate fake news detection.

Similarly, geometric deep learning frameworks [8] leverage graph structures to model complex relationships between users, content, and information sources.

Multimodal approaches have gained prominence as misinformation increasingly combines text with images and meta-data. Frameworks such as CroMe [10] and contrastive multimodal models [11] integrate visual and textual features using metric or contrastive learning strategies, achieving improved detection performance. Fu et al. [17] incorporate external knowledge and user interaction features to ground predictions in factual evidence and engagement signals. Likewise, Zhu et al. [20] propose shallow-deep multitask learning to jointly leverage unimodal and cross-modal representations, while KGAlign [18] introduces semantic-structural knowledge encoding to align textual and visual information with object-level knowledge graphs.

Knowledge integration and interpretability have also emerged as critical aspects of reliable fake news detection. Hybrid frameworks that combine machine learning with knowledge engineering enable models to validate claims against external sources and provide explainable decisions [15]. Li and Zhao [12] further augment deep learning models with vagueness detection techniques to identify ambiguously or intentionally misleading language, thereby improving transparency and trustworthiness in real-world deployments.

User privacy and personalization represent additional challenges in modern fake news detection systems. Federated Learning (FL) enables local model training on user devices while sharing only aggregated updates with a central server, thus preserving user privacy and supporting personalized recommendations [16], [17]. Although many existing systems focus primarily on binary fake-real classification, the literature demonstrates that federated learning is effective for scalable, privacy-aware news delivery [16].

Finally, human-centered factors such as critical thinking ability and media literacy significantly influence susceptibility to misinformation. Alves and Costa [9] show that individual differences in analytical reasoning and media literacy affect the effectiveness of detection systems. Fernandez et al. [16] emphasize the importance of integrating fake news detection with recommendation systems to maintain content integrity while delivering personalized user experiences. These findings support the development of hybrid frameworks that combine automated detection with user-adaptive mechanisms.

Overall, the surveyed literature reveals a clear evolution from classical machine learning approaches [5], [13], [14] to transformer-based architectures [1], [6], [7], graph and geometric learning models [3], [8], multimodal fusion techniques [10], [11], [17], [18], [20], knowledge-driven and interpretable systems [12], [15], and privacy-aware personalization strategies [16], [17]. This progression motivates the proposed hybrid framework that integrates high-capacity BERT-based fake news detection with privacy-conscious, user-centric personalization to deliver trustworthy, intelligent, and adaptive news feeds.

II. LITERATURE REVIEW

Automatic fake news detection on social media has evolved into a multidisciplinary research domain encompassing natural language processing (NLP), graph learning, multimodal representation, knowledge integration, and human-centered analysis. Early approaches primarily relied on traditional machine learning classifiers using handcrafted lexical, syntactic, and stylistic features to distinguish deceptive content from legitimate information. Techniques such as Support Vector Machines, Decision Trees, and Random Forests demonstrated reasonable performance under controlled settings; however, their dependence on manual feature engineering limited robustness across domains and rapidly evolving discourse patterns [5], [13], [14].

These limitations motivated the adoption of representation learning methods capable of capturing deeper semantic and contextual signals. Deep neural networks, particularly transformer-based language models, have significantly advanced text-based fake news detection. BERT-based architectures are widely used due to their bidirectional contextual embeddings and adaptability through fine-tuning. Ayyub et al. [6] propose a blended BERT framework that augments contextual embeddings with task-specific classification heads, improving robustness on noisy social media text. Farokhian and Rafe [7] further enhance contextual modeling by employing dual BERT encoders to capture complementary semantic perspectives, yielding improved generalization. Chian et al. [1] empirically compare traditional artificial neural networks with BERT-based models for source credibility detection, demonstrating the superior capability of transformer models in capturing subtle linguistic cues associated with deceptive content.

While transformer-based classifiers excel in semantic modeling, ensemble and hybrid architectures are frequently employed to mitigate overfitting and enhance generalization. Al-shaqi et al. [2] investigate ensemble strategies that combine diverse base learners, showing improved detection performance across heterogeneous and adversarial datasets. Comparative surveys further advocate hybrid designs that integrate interpretable handcrafted features with learned representations to balance performance and explainability [5], [13].

Beyond textual analysis, several studies exploit social and structural signals by modeling user interactions and information propagation as graphs.

Graph Neural Networks (GNNs) leverage diffusion patterns, relational dependencies, and network topology to identify coordinated misinformation campaigns and anomalous spreading behaviors [3], [8]. These approaches provide complementary evidence to textual features and are particularly effective in distinguishing organic information diffusion from orchestrated manipulation.

Multimodal misinformation detection has gained prominence due to the increasing use of images, videos, and metadata alongside text. Cross-modal transformer frameworks such as CroMe align visual and textual representations using metric or contrastive learning, improving detection accuracy when visual content supports or contradicts textual claims [10], [11]. Fu et al. [17] incorporate external knowledge sources and user interaction features into multimodal systems, demonstrating improved reliability through factual grounding. Zhu et al. [20] introduce a shallow-deep multitask learning paradigm that jointly learns unimodal and cross-modal representations, further enhancing robustness.

Knowledge integration and interpretability are increasingly recognized as critical components of reliable fake news detection systems. Knowledge graph-based methods validate claims against external facts and provide explanatory insights, thereby reducing false positives where statistical correlations alone are insufficient [15], [18]. Additionally, vagueness detection techniques identify hedging and ambiguous language patterns often used to mislead without making verifiable claims; integrating such cues with deep models enhances interpretability and assists human moderation [12].

Recent research also explores unsupervised and self-supervised learning to address the scarcity of labeled misinformation data. Structural contrastive learning frameworks exploit propagation patterns and interaction graphs to learn discriminative representations from unlabeled streams, enabling early detection of emerging misinformation topics [19]. These methods complement supervised transformer-based pipelines and support semi-supervised or continual learning scenarios.

Human-centric studies provide insights into misinformation susceptibility and mitigation strategies. Alves and Costa [9] show that critical thinking skills and new-media literacy significantly influence vulnerability to misinformation, highlighting the importance of combining automated detection with user-facing educational interventions. Fernandez et al. [16] advocate misinformation-aware recommender systems that adjust ranking and presentation strategies to reduce exposure to harmful content while preserving user engagement.

Finally, practical deployment considerations such as scalability, efficiency, and privacy are essential for real-world adoption. Recent works emphasize lightweight architectures and privacy-preserving or federated frameworks for personalization and on-device adaptation [16], [17]. Although explicit federated learning implementations are limited in this literature set, related research informs the design of systems that balance personalization with user confidentiality.

Overall, the literature reveals a clear progression: (1) a shift from feature-engineered classifiers to deep contextual models for textual understanding [1], [6], [7]; (2) integration of structural and multimodal signals through GNNs and cross-modal fusion [3], [8], [10], [11]; (3) incorporation of knowledge-based and vagueness-aware modules to improve factual grounding and interpretability [12], [15], [18]; and (4) emerging emphasis on privacy-aware and personalized detection strategies [16], [17]. Despite significant progress, most existing systems address only a subset of these dimensions, leaving a research gap for unified frameworks that simultaneously offer multimodal detection, explainability, personalization, and privacy guarantees. The present work aims to address this gap through a hybrid architecture that combines centralized high-capacity classification with lightweight, privacy-conscious personalization mechanisms.

III. COMPARISON BETWEEN MODELS

To provide a structured overview of existing approaches in fake news detection, Table I summarizes representative models reported in prior research, highlighting their feature extraction strategies and privacy considerations. This comparison facilitates a clearer understanding of prevailing trends and limitations in current systems and helps identify the research gap addressed by the proposed framework, which emphasizes accurate detection while preserving user privacy.

Table I presents a comparative analysis of commonly used fake news detection models, including Artificial Neural Networks (ANNs), BERT-based architectures, Graph Neural Networks (GNNs), and multimodal frameworks. Traditional ANN-based models primarily rely on handcrafted textual features and demonstrate limited adaptability across domains. Transformer-based models such as BERT significantly improve detection accuracy by leveraging deep contextual embeddings; however, they are typically trained in centralized settings that require access to large volumes of user-generated data, raising privacy concerns. Similarly, GNN-based approaches effectively capture relational and propagation patterns within social networks but often depend on centralized aggregation of interaction data, which can expose sensitive user information.

Recent studies have begun exploring federated deep learning frameworks to address privacy challenges by enabling collaborative model training without sharing raw user data. While these approaches improve privacy preservation, they often focus on model training efficiency and may not fully exploit high-capacity language models for fine-grained content understanding.

The proposed hybrid system advances the state of the art by integrating a robust BERT-based fake news classifier with federated, on-device personalization mechanisms. This design combines the strong semantic modeling capability of transformer architectures with privacy-preserving distributed learning, ensuring effective fake news detection while safeguarding sensitive user data. As a result, the proposed approach represents a comprehensive step toward privacy-aware, trustworthy, and personalized content recommendation systems.

TABLE I
COMPARISON OF DIFFERENT FAKE NEWS DETECTION MODELS

Model	Feature Extraction
ANN & BERT Hybrid	Contextual and source credibility embeddings
Blended BERT Model	Linguistic and semantic contextual embeddings
Dual BERT Network	Parallel contextual representations
GNN	User-post relational graph embeddings
DNN	Deep semantic and lexical features

IV. CONCLUSION

This survey highlights the rapid evolution of fake news detection research, tracing its progression from traditional feature-engineered approaches to advanced deep learning, multimodal, and graph-based architectures. Early studies primarily employed classical machine learning algorithms such as Support Vector Machines, Decision Trees, and Random Forests, relying heavily on handcrafted linguistic and stylistic features. Although these methods achieved initial success, their limited contextual understanding and poor cross-domain generalization motivated the shift toward representation learning techniques.

The introduction of transformer-based models, particularly BERT and its variants, marked a significant advancement in fake news detection by enabling deeper contextual and semantic modeling. Works such as Ayyub et al. [6] and Farokhian and Rafe [7] demonstrate that contextual embeddings substantially improve robustness and adaptability when dealing with noisy and informal social media text. Ensemble learning strategies and hybrid pipelines further enhance generalization and resilience against domain shifts.

Recent research also emphasizes the importance of incorporating multiple data modalities and relational structures. Multimodal frameworks [10], [11], [17], [20] integrate textual, visual, and contextual information to capture cross-modal inconsistencies and correlations, while graph neural networks and geometric deep learning approaches [3], [8] exploit social interaction and propagation patterns to detect coordinated misinformation campaigns. These findings indicate that effective misinformation detection requires understanding both content and its social dissemination dynamics.

In parallel, interpretability and factual grounding have emerged as key priorities. Knowledge-based methods leveraging external knowledge graphs [15], [18], along with vagueness detection techniques [12], enhance transparency and reliability by aligning model decisions with verifiable facts and explainable linguistic cues. Additionally, self-supervised and contrastive learning approaches [19] address label scarcity by enabling models to learn discriminative representations from unlabeled data streams.

Human-centered and ethical considerations further enrich this research landscape. Studies on media literacy and user behavior [9] underline that misinformation mitigation is not solely a technical challenge but also a socio-cognitive one. Misinformation-aware recommender systems and privacy-preserving personalization frameworks [16], [17] reflect a growing effort to balance detection accuracy, user engagement, and ethical data usage.

Overall, the surveyed literature reveals a clear trajectory toward integrated solutions. Future fake news detection systems must unify deep contextual modeling, multimodal fusion, graph-based reasoning, and privacy-aware personalization to deliver scalable, interpretable, and user-centric solutions. The field is steadily advancing toward comprehensive frameworks that balance accuracy, transparency, and ethical responsibility in combating misinformation.

REFERENCES

- [1] T. H. C. Chiang, C. S. Liao, and W. C. Wang, "Investigating the difference of fake news source credibility recognition between ANN and BERT," *Journal of Information Technology and Media Studies*, vol. 9, no. 2, pp. 45–52, 2022.
- [2] M. Al-shaqi, D. B. Rawat, and C. Liu, "Ensemble techniques for robust fake news detection," in *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pp. 1–8, 2024.
- [3] Y. Han, S. Karunasekera, and C. Leckie, "Graph neural networks with continual learning for fake news detection from social media," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 2046–2059, 2020.
- [4] G. Guler and S. Demirci, "Deep learning based fake news detection on social media," *International Journal of Computer Applications*, vol. 180, no. 34, pp. 1–7, 2023.
- [5] D. Singh, H. Panwar, and S. Chaturvedi, "Designing a fake news detection system using machine learning techniques," *International Journal of Advanced Research in Computer Science*, vol. 16, no. 5, pp. 102–108, 2025.
- [6] S. Ayyub, A. Khan, and R. Malik, "BERT-based blended approach for fake news detection," *Journal of Computational Linguistics Research*, vol. 15, no. 3, pp. 215–223, 2021.
- [7] M. Farokhian and V. Rafe, "Fake news detection using dual BERT deep neural networks," *Neural Computing and Applications*, vol. 35, no. 9, pp. 7511–7520, 2023.
- [8] L. Zhang, Y. Liu, and P. Chen, "Fake news detection on social media using geometric deep learning," *Expert Systems with Applications*, vol. 237, pp. 121–130, 2024.
- [9] N. Alves and M. Costa, "Fake news detection on social media: The predictive role of university students' critical thinking dispositions and new media literacy," *Computers in Human Behavior*, vol. 136, 2022.
- [10] S. Patel, A. Nair, and J. Lee, "CroMe: Multimodal fake news detection using cross-modal tri-transformer and metric learning," *IEEE Access*, vol. 11, pp. 134056–134069, 2023.
- [11] R. Kumar and V. Menon, "Multimodal fake news detection with contrastive learning and optimal transport," *Information Fusion*, vol. 103, pp. 18–29, 2024.
- [12] H. Li and X. Zhao, "Combining vagueness detection with deep learning to identify fake news," *Journal of Artificial Intelligence Research*, vol. 76, pp. 310–326, 2023.
- [13] P. Reddy and M. Thomas, "A comparative study of machine learning and deep learning techniques for fake news detection," *Procedia Computer Science*, vol. 197, pp. 512–520, 2022.
- [14] D. Bhattacharya and S. Kumar, "Detection of fake news using machine learning and natural language processing algorithms," *Applied Intelligence*, vol. 54, pp. 4205–4219, 2024.
- [15] E. Garza and M. Rodríguez, "Combining machine learning with knowledge engineering to detect fake news in social networks—A survey," *Knowledge-Based Systems*, vol. 233, pp. 107–125, 2021.
- [16] A. Fernandez, S. Iqbal, and R. Green, "Advancing Misinformation Awareness in Recommender Systems for Social Media Information Integrity," *Information Processing and Management*, vol. 62, no. 4, pp. 101–115, 2025.
- [17] L. Fu, J. Zhang, H. Li, and X. Wang, "Multimodal Fake News Detection Incorporating External Knowledge and User Interaction Feature," *Wiley Online Library*, 2023.
- [18] T.-V. La, M.-H. Nguyen, and M.-S. Dao, "Fake news detection," *arXiv preprint*, 2025.
- [19] X. Chen, L. Zhou, and Y. Li, "An unsupervised fake news detection framework based on structural contrastive learning," *Cybersecurity*, vol. 8, no. 1, 2025.
- [20] Y. Zhu, Y. Wang, and Z. Yu, "Multimodal fake news detection: MFND dataset and shallow-deep multitask learning," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 891–899, 2025.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)