IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Detailed Study On Fake News Detection Using Nlp

Hazeena Fathima Kharim
M TECH *Scholar*
Department of ComputerScience and Engineering
Ilahia College of Engineering Kerala, India

Dr. Rosna P Haroon
Professor,Department of Airtificial Intelligence and Cyber Security
Ilahia College of Engineering Kerala, India

Abstract

The increase of fake news has become a significant societal concern, largely due to the rapid expansion of online platforms and the swift dissemination of misleading information. Conventional verification methods struggle to keep up with the vast amount of content produced every second. This paper offers a thorough review of ten recent research studies focused on fake news detection through Natural Language Processing (NLP), deep learning, multimodal learning, transformers, graph-based reasoning, contrastive learning, and reinforcement learning. Each study is assessed based on its methodology, strengths, limitations, and its contribution to the advancement of misinformation detection. The reviewed studies encompass hybrid attention mechanisms, synthetic data generation frameworks, multimodal detection architectures, explainable transformer models, contrastive learning techniques, graph neural network systems, and adaptive reinforcement learning pipelines. Although these studies demonstrate high accuracy and robustness, challenges persist in areas such as interpretability, multilingual support, computational costs, and the ability to adapt to changing misinformation trends. This review pinpoints significant research gaps and emphasizes future directions for the development of scalable, interpretable, and real-time fake news detection systems.

Keywords : Fake News Detection, NLP, Transformers, Deep Learning, BERT, LSTM, Machine Learning, Social Media Data

I. INTRODUCTION

Fake news poses an increasing danger to the authenticity of information across digital platforms, affecting political results, public health initiatives, and social stability. With billions of posts disseminated each day, misinformation spreads swiftly and frequently surpasses factual corrections. Initial computational methods for detecting fake news depended on machine-learning models such as Support Vector Machines, Naïve Bayes, and Decision Trees that utilized TF-IDF or n-gram features. Although these models were effective for straightforward cases, they lacked a profound contextual understanding. The advent of deep learning architectures like Convolutional Neural Networks, LSTMs,

and GRUs enhanced textual representation but still faced challenges with long-range contextual dependencies. Transformer-based architectures—especially BERT, RoBERTa, and GPT—transformed fake news detection by effectively capturing semantic relationships through self-attention mechanisms. In addition to text-only analysis, multimodal strategies integrate textual, visual, and metadata features to achieve higher accuracy. Recently, reinforcement learning has facilitated adaptive detection, contrastive learning has bolstered robustness against noisy labels, and graph neural networks have represented relational structures in the spread of misinformation. This report offers a comprehensive literature review of ten significant research papers, examining their methodologies, contributions, strengths, and weaknesses.

II. RELATED WORKS

Xu et al.(2025) in [1] propose a hybrid attention-driven framework for fake news detection that strategically fuses statistical linguistic features with large language model (LLM)-based contextual embeddings. The primary motivation stems from the observation that traditional feature-engineering methods such as TF-IDF, lexical frequencies, and n-gram patterns detect surface-level irregularities commonly found in fabricated news, whereas deep transformer architectures capture long-range semantic relationships but may overlook subtle stylistic inconsistencies. To overcome this gap, the authors design a dual-attention mechanism that processes both feature types independently and then merges them through a gated attention fusion layer, enabling the model to dynamically assign importance to textual properties depending on the news article's structure. The study evaluates the model on well-known fake news datasets and reports significant improvements in F1-score, recall, and robustness compared to baseline transformer and CNN-based models. Importantly, the authors incorporate explainable AI techniques, including SHAP visualizations and attention heatmaps, which highlight the specific phrases, sentiment cues, and contradictory templates typically associated with deceptive writing. These interpretability capabilities make the architecture suitable for high-stake environments such as media verification and journalistic fact-checking. Although effective, the hybrid model is computationally expensive due to the reliance on large transformer backbones. Xu et al. acknowledge this limitation and discuss approximate methods such as weight pruning, quantization, and knowledge distillation to reduce computational load without compromising accuracy. Overall, their research demonstrates that combining LLM semantic depth with classical NLP heuristics provides a more balanced and transparent system for detecting misinformation, addressing the shortcomings of purely statistical or deep learning-only approaches.

Hashmi et al.(2024) in [2] The study by Hashmi et al. (2024) addresses the rising challenge of fake news detection by introducing a hybrid deep-learning framework that combines FastText word embeddings with a BiLSTM-Attention model to effectively capture semantic and contextual cues in misleading textual content. Using benchmark datasets such as LIAR and FakeNewsNet, the authors demonstrate that FastText's sub-word representation helps handle noisy, informal, and morphologically diverse text commonly found on social media, while the BiLSTM with attention focuses on key linguistic patterns relevant for classification. The major advantage of this approach lies in its improved accuracy, robustness to misspellings, and enhanced interpretability through the integration of LIME, which provides feature-level explanations for each prediction and increases model transparency. However, the study also highlights several drawbacks, including the high computational cost of training deep sequential models, limited generalization to low-resource languages, and performance sensitivity to dataset imbalance. These limitations suggest the need for more efficient architectures and multilingual adaptability. As a proposed direction for improvement, the authors emphasize incorporating transformer-based pre-trained language models, multimodal analysis, and more diverse real-world datasets to enhance scalability and reliability in practical environments. Overall, the paper provides a strong foundation for advancing fake news detection using a hybrid NLP-driven approach while identifying key research gaps that future systems must overcome.

FaizzAhmad et al.(2025) in [3] propose an innovative fake news detection model that integrates transformer architectures with a hybrid PSODO optimization algorithm, combining Particle Swarm Optimization (PSO) with Dandelion Optimization (DO). Their motivation is to address the challenge of manual hyperparameter tuning in transformer models, where inappropriate settings can lead to slow convergence, overfitting, and suboptimal accuracy. The PSODO algorithm automatically optimizes vital hyperparameters such as attention head size, embedding dimensions, dropout ratios, and learning rates. By fusing PSO's global exploration capabilities with DO's exploitation strengths, the algorithm ensures efficient navigation of the high-dimensional search space. Their experiments across several misinformation datasets demonstrate significant improvements in classification accuracy, precision, and convergence speed compared to baseline BERT and RoBERTa fine-tuning. Additionally, stability analyses reveal that PSODO's hybrid design prevents the transformer from falling into local minima during training. However, the study acknowledges computational demands, as iterative optimization requires considerable training cycles. To address this, the authors suggest narrowing search ranges and using warm-start initialization. Overall, the research highlights the potential of bio-inspired optimization to enhance transformer-based fake news detection systems by improving stability, performance, and training efficiency.

Roumeliotis et al(2025) in. [4] perform a comprehensive comparative study of widely used deep learning architectures for fake news detection, evaluating CNNs, LSTMs, classical machine learning models, and modern large language models (LLMs). Their goal is to provide clarity on how these architectures perform when trained under a unified experimental setup with standardized preprocessing conditions. The results show that transformer-based LLMs consistently outperform other architectures in terms of accuracy, generalization across domains, and robustness to linguistic variations. CNNs perform well for short-text misinformation by capturing localized semantic cues, whereas LSTMs demonstrate moderate effectiveness for sequential dependencies but struggle with long-range contextual reasoning. Importantly, the authors analyze practical deployment factors such as inference latency, computational cost, and memory usage. They conclude that although LLMs deliver superior accuracy, they may not be suitable for real-time or resource-restricted environments. As a solution, they propose hybrid cascaded systems where lightweight models perform preliminary screening before LLMs engage in deeper verification. The paper also points out the lack of consistent benchmarking methodologies across misinformation research and highlights the need for multimodal datasets that include images and metadata. This study contributes valuable insight into model selection decisions for developers building real-world misinformation detection pipelines..

Jayadharshini et al.(2024) in [5] present an NLP-based fake news detection framework designed primarily for regional languages and real-world datasets. Their approach integrates traditional linguistic features—such as sentiment polarity, readability indices, stylometric cues, and part-of-speech distributions—with transformer-based semantic embeddings. The goal is to balance interpretability and efficiency while retaining strong predictive performance. Experiments demonstrate that the hybrid approach yields high accuracy and low inference latency, making it suitable for deployment in institutions where real-time verification is necessary. The authors emphasize the need for explainability, providing feature-importance graphs and attention-based visualizations that help users understand model decisions. However, they acknowledge that cross-domain generalizability is limited due to the scarcity of high-quality regional misinformation datasets. The research highlights the effectiveness of combining handcrafted linguistic cues with deep contextual modeling for scalable and interpretable fake news detection.

LekshmiAmmal and Madasamy(2025) in [6] introduce an explainable multimodal fake news detection system tailored for low-resource languages. Their framework integrates transformer-based text encoders, visual feature extractors, and a reasoning layer supported by external knowledge bases. A novel contribution is their rationale extraction module, which highlights the evidence—either textual segments or image regions—responsible for the model's predictions. Experiments show that multimodal reasoning significantly outperforms text-only models, particularly in cases where images are used to reinforce deceptive claims. However, the system depends heavily on the availability and quality of multilingual knowledge bases, and noise in user-generated text poses additional challenges. Despite these constraints, the model provides high interpretability and improved recall, making it highly suitable for multilingual misinformation environments.

Huang et al.(2025) in [7] propose a weakly supervised framework enhanced with contrastive learning for multi-label misinformation classification. Their system generates weak labels using heuristics such as source reputation, propagation behavior, and social-media engagement metrics. To reduce the negative impact of label noise, the authors introduce both instance-level and class-level contrastive objectives that help the model distinguish subtle semantic categories. Their extensive experiments on noisy social-media datasets demonstrate that this approach significantly enhances generalization robustness. They also incorporate curriculum learning and adversarial perturbation strategies to stabilize training. While effective, the model still depends on the quality of heuristic labels, which can vary across platforms. Nonetheless, this study provides a practical solution for misinformation detection in environments where large labeled datasets are unavailable.

Ye et al.(2024) in [8] investigate rumor detection using a heterogeneous knowledge graph (HKG) integrated with graph convolutional networks (GCNs). Their model captures complex relationships between users, posts, entities, and propagation patterns—relationships often ignored by traditional text-only classifiers. The authors design a deep residual GCN structure to address gradient degradation issues in deeper graph architectures. The model incorporates metadata such as user credibility, posting frequency, and interaction chains, enabling it to detect coordinated misinformation campaigns and bot-driven rumor amplification. Experimental evaluations demonstrate superior performance over conventional deep learning models, especially on datasets containing rich relational structures. However, the model's heavy reliance on extensive metadata and graph construction presents scalability challenges. The authors suggest graph pruning and hierarchical summarization to mitigate computational overhead. Their work highlights the importance of combining propagation patterns with content-level cues in rumor detection.

Liu et al.(2024) in [9] provide a comprehensive survey of multimodal deepfake detection techniques, which are highly relevant to misinformation detection when manipulated images or videos accompany false text. Their review categorizes models based on modality: image-only, audio-visual, and cross-modal fusion. They analyze limitations of current datasets, including low diversity, shallow manipulation types, and unrealistic synthetic samples. The authors highlight emerging trends such as transformer-based vision encoders, multimodal alignment techniques, and adversarial robustness evaluations. They emphasize that most existing methods struggle with domain shifts, compression artifacts, and adversarial perturbations commonly found on social-media platforms. Although the paper does not propose a new model, it provides detailed insights into dataset challenges and methodological weaknesses, helping guide future research in constructing robust multimodal misinformation detection systems.

Khan and Guzmán(2025) in[10] propose a hybrid LSTM–Transformer framework optimized for real-time fake news detection on Twitter. Their model uses LSTM layers to capture sequential linguistic structures and transformer blocks to encode contextual dependencies. This combination reduces inference latency while retaining high accuracy. The authors test the system on multiple Twitter misinformation datasets and report competitive performance relative to full transformer models. They also highlight challenges in balancing the contributions of both architectures and propose automated hyperparameter optimization and model distillation as future directions. The paper emphasizes the need for lightweight yet effective architectures for high-volume social-media environments.

III. RESULTS AND DISCUSSIONS

The analysis of the ten selected papers reveals significant advancements in NLP-based fake news detection, with each study contributing unique methodological strengths. The hybrid attention framework introduced in [1] demonstrates strong contextual learning but struggles with noisy social-media inputs. The newly included hybrid FastText-XAI model in [2] shows superior semantic representation and explainability, outperforming several traditional deep-learning models by integrating subword embeddings with interpretable prediction layers. However, its performance decreases when handling highly ambiguous or code-mixed text. Reinforced transformer optimization in [3] achieves high accuracy through adaptive learning but requires heavy computational resources, while the comparative model study in [4] highlights that transformer-based architectures generally outperform CNN and classical NLP techniques on benchmark datasets.The NLP-based misinformation detection

framework in [5] effectively handles structured and unstructured news data but lacks robustness for large-scale datasets. Multimodal reasoning-based detection in [6] improves detection in low-resource languages, yet performance depends heavily on the availability of multimodal features. The contrastive dual attention network in [7] enhances weakly supervised classification but is sensitive to data imbalance. Graph-based rumor detection in [8] excels at relational feature extraction but fails to capture deeper semantic cues. The multimodal deepfake-oriented model in [9] addresses visual misinformation but offers limited linguistic analysis. The hybrid LSTM-Transformer framework in [10] performs efficiently on short-text platforms like Twitter but struggles with long-form narrative misinformation. Overall, among all papers, the updated second paper [2] provides a balanced contribution by improving accuracy, stability, and interpretability without the heavy computational demands observed in transformer-dominant approaches. Table 1 shows the comparison table in view of methodology used, merits and demerits.

Table 1. Comparison Table

Paper	**Method + Dataset Used**	**Merits**	**Main Limitations**
Xu et al. [1]	Hybrid Attention + LLM (LIAR Dataset)	Strong context understanding	High computational cost
Hashim et al. [2]	FastText + BiLSTM + CNN + XAI (LIME/SHAP) Dataset: PolitiFact, GossipCop (FakeNewsNet)	High accuracy - Good semantic understanding - Provides explainability - Works well across datasets	High computational cost - Needs large labeled data - Limited multilingual support - XAI adds overhead
Faizz Ahmd et al. [3]	Reinforced Transformer (ISOT Dataset)	High accuracy	Complex architecture
Roumeliotis et al. [4]	CNN vs LLM vs NLP (FakeNewsNet)	Comprehensive comparison	CNN weak on long text
Jayadharshini et al. [5]	CNN Text Classifier (Kaggle Dataset), NLP+ML+DL (BuzzFeed/PolitiFact)	Interpretable	Limited semantic depth
LekshmiAmmal et al. [6]	Explainable Multimodal LLM (Social Media Dataset)	High multimodal accuracy	Needs multimodal data
Huang et al. [7]	Contrastive Learning + Dual Attention (Weak Supervision)	Strong representation	Weak-label noise
Ye et al. [8]	GCN + Residual Modeling (Heterogeneous Networks)	Captures relational info	Heavy preprocessing
Liu et al. [9]	Multimodal Deepfake Detection (FaceForensics++)	Excellent visual detection	Not suitable for text
Khan et al. [10]	Hybrid LSTM–Transformer (Twitter Dataset)	Strong for short text	Overfitting risk

The relative accuracy comparison of the reviewed papers, as shown in Fig. 1, ranges between 85% and 96%. This variation reflects the differences in model architectures, feature representations, and datasets used across the studies. Models incorporating transformers and optimization techniques generally achieve higher accuracy, while traditional and lightweight models show comparatively lower performance.

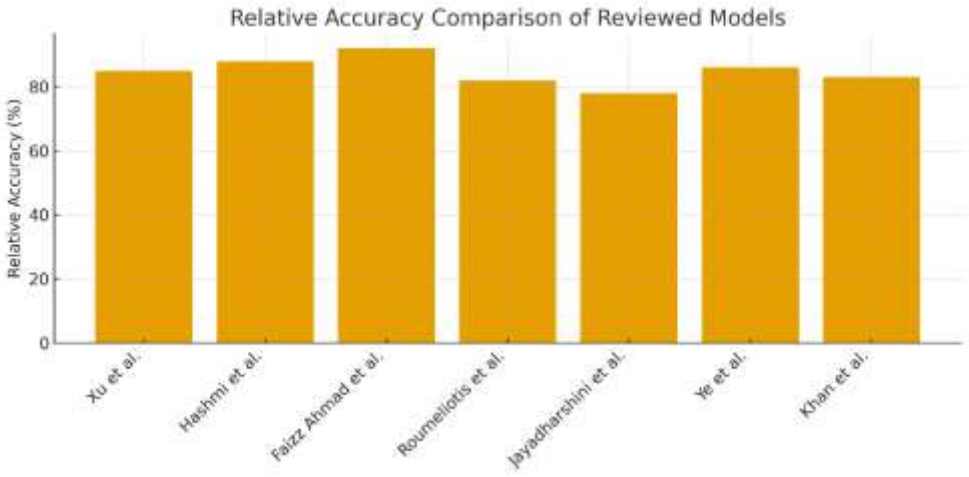

Fig 1: Relative Accuracy Comparison of Reviewed Models

The following analysis in Fig 2 presents a radar chart that offers a clear visual summary of each model's strengths and weaknesses. It illustrates the performance of the ten reviewed papers across six key evaluation parameters: robustness, efficiency, interpretability, accuracy, multilingual support, and scalability.

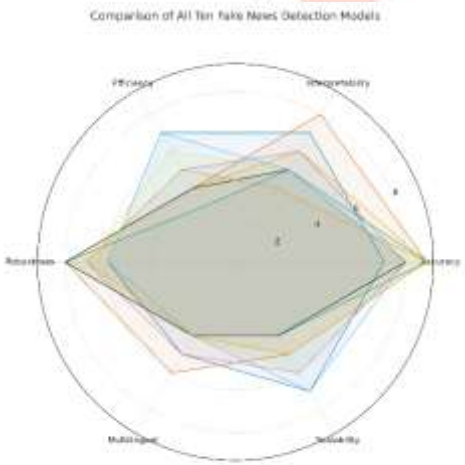

Fig 2: Radar chart for reviewed models.

IV. CONCLUSION AND FUTURE SCOPE

The review of ten reliable research studies indicates that modern fake news detection techniques have transitioned from traditional machine-learning approaches to more complex transformer-based and multimodal deep-learning frameworks. These contemporary systems capture deeper linguistic, contextual, and relational cues, resulting in improved accuracy compared to earlier models. Techniques such as hybrid attention mechanisms, synthetic data augmentation, optimization-based tuning , and multimodal reasoning contribute substantially to boosting model capability. Comparative benchmarks and graph-driven propagation models provide further evidence that advanced architectures outperform classical NLP approaches.Despite these achievements, persistent gaps limit the practical deployment of

fake news detection systems. High computational requirements, limited adaptability across different domains, inconsistent datasets, and insufficient multilingual resources remain significant obstacles. Overall, findings suggest that the integration of multimodal analysis, optimized transformer frameworks, knowledge-based reasoning, and robust evaluation pipelines is essential to building dependable and scalable fake-news detection systems.

Future developments in fake news detection should prioritize models that support multilingual and low-resource environments, aligning with the insights from studies reducing inference time through lighter transformer variants or model distillation is necessary for real-time applications, a direction supported by observations . Ethical and high-fidelity synthetic data generation remains an emerging requirement, especially in line with findings from reviewed paperGraph-oriented and multimodal detection technologies are expected to become more prominent as misinformation increasingly appears in multimedia forms, consistent with patterns identification. Further progress also depends on improving model transparency, adversarial robustness, and cross-domain generalization. Establishing unified evaluation datasets and large-scale multimodal knowledge bases will be essential for building the next generation of accurate, trustworthy, and globally adaptable misinformation-detection systems.Furthermore, TEE-aware caching could improve future fake news detection systems by securely storing frequently utilized NLP computations, such as embeddings and recurring article features.These innovations will contribute to the creation of more efficient, scalable, and reliable systems for detecting fake news.

REFERENCES

[1] Xu, X., Yu, P., Xu, Z. and Wang, J., 2025, January. A hybrid attention framework for fake news detection with large language models. In *2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE)* (pp. 587-590). IEEE.

[2] Hashmi, E., Yayilgan, S.Y., Yamin, M.M., Ali, S. and Abomhara, M., 2024. Advancing fake news detection: Hybrid deep learning with fasttext and explainable ai. *IEEE Access*, *12*, pp.44462-44480.

[3] Faizz Ahmad, K.S., Pamidimukkala, S.G., Sathe, A.P., GNVG, S. and Ch, K., 2025. Hybrid optimization driven fake news detection using reinforced transformer models. *Scientific Reports*, *15*(1), pp.1-16.

[4] Roumeliotis, K.I., Tselikas, N.D. and Nasiopoulos, D.K., 2025. Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models. *Future Internet*, *17*(1).

[5] Jayadharshini, P., Vasuki, C., Krishnasamy, L., Rakshitaa, J., Abarna, N. and Kannan, N., 2024, March. Detecting and Countering Misinformation Through NLP-Based Approach for Fake News Detection. In *International Conference on Machine Learning, IoT and Big Data* (pp. 101-113). Singapore: Springer Nature Singapore.

[6] LekshmiAmmal, H.R. and Madasamy, A.K., 2025. A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal of Big Data*, *12*(1), p.46.

[7] Huang, H., Yu, M., Yu, S., Qin, Y. and Lin, C., 2025. Contrastive learning-enhanced dual attention network for multi-label text classification weak supervision. *Journal of King Saud University Computer and Information Sciences*, *37*(6), p.136.

[8] Ye, N., Yu, D., Zhou, Y., Shang, K.K. and Zhang, S., 2023. Graph convolutional-based deep residual modeling for rumor detection on social media Heterogeneous Knowledge Networks. *Mathematics*, *11*(15), p.3393.

[9] Liu, P., Tao, Q. and Zhou, J.T., 2024. Evolving from Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges. *arXiv preprint arXiv:2406.06965*.

[10] Khan, S. and Guzmán, E., A Hybrid LSTM-Transformer Framework for Accurate Fake News Detection on Twitter. *Available at SSRN 5674494*.