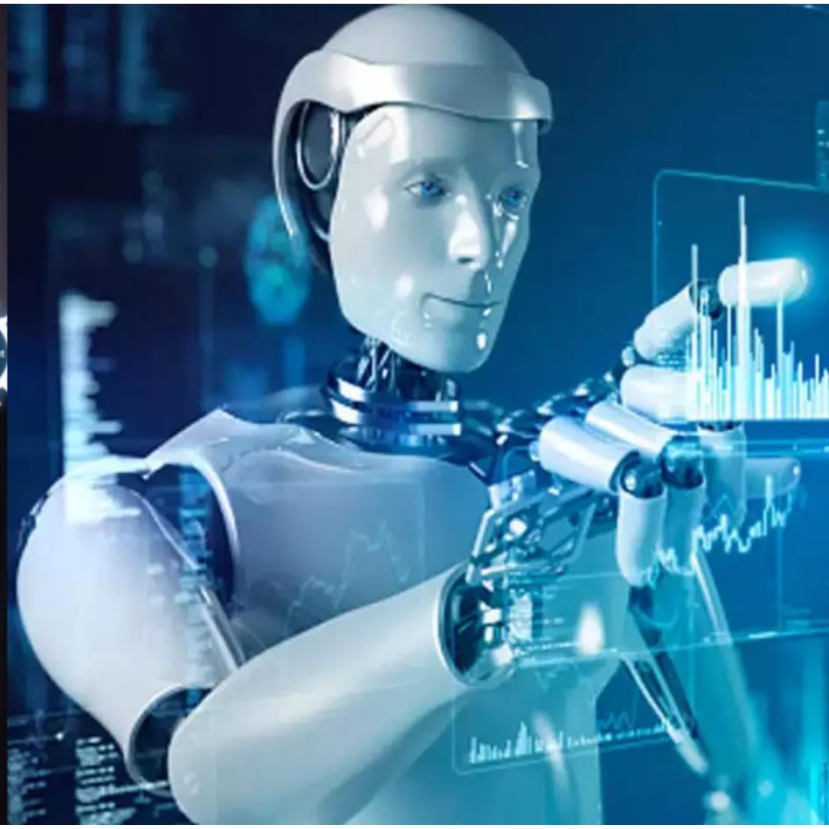
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 10, October 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

Fake News Detection in Social Media

Sanket Mohan Kotkar, Soudamini T. Somvanshi

Department of Computer Engineering, Dr. D.Y. Patil College of Engineering, Akurdi, Pune, Maharashtra, India

Department of Computer Engineering, Dr. D.Y. Patil College of Engineering, Akurdi, Pune, Maharashtra, India

ABSTRACT: The rise of social media has changed the way data is shared worldwide but at the same time, it has promoted the spread of lies at a much higher rate. The rising problem of misinformation has impacted different areas like politics, medicine, the market, and the level of trust in society and this fact has made human moderation to be of no use. First methods for fake news recognition by using conventional machine learning techniques like Logistic Regression, Naïve Bayes, and Support Vector Machines revealed that these methods do not understand the semantics deeply. Yet, the subsequent development of deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) has improved both the contextual and sequential understanding. The recent transformer-based models, especially Large Language Models (LLMs) like BERT, RoBERTa, and GPT-4, have gone beyond existing limits by enabling complex semantic reasoning and being inherently multilingual.

This review traces the landmark changes in fake news detection from 2016 to 2025, with a focus on datasets, methods, and performance metrics. It acknowledges the issues of dataset biases, the necessity for real-time scalable solutions, and the problem of explainability among its challenges. In addition, it offers a set of potential research directions that revolve around multimodal detection, graph-based learning, and Explainable Artificial Intelligence (XAI).

KEYWORDS: Detection of Misinformation, Social Networking Platforms, Artificial Intelligence, Machine Learning Methods, Deep Learning Models, Transformer Architecture, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) Methodologies, Graph Neural Network (GNN) Frameworks, Multimodal Detection Strategies, Explainable Artificial Intelligence (XAI)

I. INTRODUCTION

The social networks Facebook, Twitter, WhatsApp, and YouTube became indispensable channels for information dissemination all around the world. Due to their user-friendly interface and interactive functionality, information is spread rapidly, but they also lead to the uncontrollable spreading of fallacies. Misinformation, as incorrectly misleading content disguised as reliable news, is a danger to democratic regimes, a breach of public faith, and an interference with social cohesion [1] [2].

Unlike traditional news that is governed by editorial standards, social networks operate with no centralized power, such that bad actors can spread misinformation at high speeds [3]. High-profile events including the COVID-19 pandemic, the 2020 Delhi riots, as well as the 2016 American elections, suggest the significant real-world effects of misinformation [4]. It is not practical to moderate the sheer volume and velocity of online content with human moderation alone. As a consequence, computationally driven solutions rooted in artificial intelligence have become inevitable [5] [6].

This paper is an in-depth survey of deep learning techniques for detecting fake news, dataset construction, and performance requirements between 2016 and 2025. It also provides some thoughts regarding the trend for the future based on problems that were faced as well as advancements in large language models (LLMs), graph neural networks (GNNs), and retrieval-augmented generation (RAG) systems [7] [8].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

II. NEED FOR AUTOMATED DETECTION

Automated identification of misinformation is essential for various reasons:

- **Volume of Data:** There are millions of posts created every minute, which surpasses the ability of humans to moderate effectively.
- **Rapid Spread:** Misinformation can circulate more quickly than humans can verify facts, particularly during emergencies or electoral events.
- **Human Constraints**: The process of manually identifying fake news is often slow, expensive, and susceptible to biases.
- **Capacity for Scaling:** Automated solutions can monitor large networks in real time efficiently.
- **Multimodal & Multilingual Coverage:** AI technologies are capable of analyzing text, visual content, and video material in various languages.
- **Consistency:** Automation ensures uniform and unbiased decisions across different contexts.

III. EVOLUTION OF FAKE NEWS DETECTION TECHNIQUES

Fake news detection was pioneered with classical machine learning, from transformer-based LLMs, to being augmented by graph-based, retrieval-augmented models. Those models increase accuracy, scalability, and multilinguality [1], [2].

A. Conventional Machine Learning (2016–2017)

The initial systems were linear machine learning models, i.e., Logistic Regression, Naïve Bayes, Support Vector Machines (SVMs), Random Forests, and K-Nearest Neighbors (KNNs) [1]. The models were implemented with hand-crafted text representations, i.e., Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and N-grams. While these models are predictive as well as interpretable, they do not learn deeper semantic relationships among words [2].

Accuracy Achieved: 70–85% (Datasets: ISOT, Kaggle Fake News)

B. Deep Learning Approaches (2018–2019)

To overcome the feature limitation, the authors explored deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM) networks [2][3]. Semantic representation was enhanced through the use of word embeddings, i.e., Word2Vec and GloVe. CNNs were suitable for the local features, whereas LSTMs were used to capture long-term dependencies, thus the accuracy was higher than that of traditional methods [4].

Accuracy Achieved: 85–92% (Datasets: Twitter15, Twitter16, Weibo)

C. Hybrid Deep Learning Models (2020–2021)

The merged design used a combination of CNN and LSTM/GRU layers, enhanced with attention mechanisms to efficiently identify the most relevant textual signals [5]. Such a method was capable to handle local as well as sequential dependencies, thus it was able to raise considerably the correctness of the recognition of event-driven or viral misinformation.

Accuracy Achieved: 90–95% (Datasets: FakeNewsNet, PHEME, ISOT)

D. Transformer-Based Models (2022–2023)

Basically, the whole area of detecting fake news has been changed drastically by the arrival of transformer architectures. The above-mentioned models such as BERT, RoBERTa, XLNet, ALBERT, and DistilBERT are equipped with self-attention modules that can efficiently deal with long-range dependencies [6][7]. The impressive accuracies are essentially the outcomes of doing fine-tuning of these pre-trained models with domain-specific data.

Accuracy Achieved: 92–97% (Datasets: CoAID, COVID-HeRA, FakeNewsNet, Weibo)

E. Advanced LLMs and Graph-Based Methods (2024–2025)

Current advancements combine Graph Neural Networks (GNNs) and Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) such as ChatGPT, GPT-4, LLaMA, and Gemini [8]. GNNs help in understanding the spreading of falsified information, and RAG is used for fetching the latest facts from reliable sources to verify the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

statements made. The interaction of these different technologies results in the effectiveness levels which are close to the human ones and still has the potential to be better through Explainable AI (XAI) co which is targeted to enhance the interpretability aspect [8].

Accuracy Achieved: 97–99.2% (Datasets: CheckThat-2022, MediaEval, FakeNewsNet)

Evolution of Fake News Detection Techniques (2016–2025)

Year Range	Techniques / Technology Used	Datasets Used (Examples)	Accuracy Achieved	Key Features / Remarks
2016–2017	Logistic Regression, Naïve Bayes, SVM, Random Forest, KNN; BoW, TF-IDF, N-grams	ISOT, Kaggle Fake News	70–85%	Simple, interpretable models; limited context understanding; relied on handcrafted features.
2018–2019	Deep Learning (CNNs, LSTMs, Bi-LSTMs); Word2Vec, GloVe embeddings	Twitter15, Twitter16, Weibo	85–92%	Captured semantic & sequential meaning; reduced reliance on manual features.
2020–2021	Hybrid Models (CNN+LSTM, Bi-LSTM+GRU, Attention Mechanism, FastText)	FakeNewsNet, PHEME, ISOT	90–95%	Combined local + sequential features; attention improved focus on key words/phrases.
2022–2023	Transformer Models (BERT, RoBERTa, XLNet, ALBERT, DistilBERT, multimodal VGG19)	CoAID, COVID-HeRA, FakeNewsNet, Weibo	92–97%	Self-attention captured long-range dependencies; domain-specific fine-tuning improved results.
2024–2025	Large Language Models (GPT-4, LLaMA, Gemini), Graph Neural Networks (GNNs), RAG, FakeBERT	CheckThat-2022, MediaEval, FakeNewsNet, Twitter15/16	97–99.2%	Multilingual, multimodal, explainable AI; near-human accuracy with fact verification.

IV. LATEST ADVANCEMENTS IN FAKE NEWS DETECTION (2024–2025)

The recent changes in methods for detecting false information comprise of advanced techniques such as Retrieval-Augmented Generation (RAG), Graph Neural Networks (GNNs), and Large Language Models (LLMs). These methods, as opposed to classic classification models, combine content analysis, propagation learning, and checking with external factual sources. They are therefore able to reach performance levels on carefully calibrated benchmark datasets that are very close to those of human evaluators, with the accuracy going from 97% to as high as 99.2%.

A. Large Language Models (LLM's)

Such models as Gemini, LLaMA, and GPT-4 are using billions of parameters to deepen their contextual understanding. This feature is the main reason why they demonstrate outstanding results in zero-shot and few-shot settings and at the same time they are capable of efficient multilingual recognition in different domains. In fact, they can be sarcastic, biased, or deceived without losing their focus. The most important mechanism in transformer-based models is that of self-attention, and it can mathematically be represented as

:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

B. Retrieval-Augmented Generation(RAG)

Retrieval-Augmented Generation strengthens LLMs by retrieving from trusted external knowledge, for example, knowledge graphs or validated databases, before generating responses. This helps to strongly decrease hallucinations while enhancing factual coherence in tests, such that it could probabilistically model as:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

$$P(y|x) = \sum_{z} P(y|x,z)P(z|x)$$

where z is the retrieved evidence.

C. Graph Neural Networks (GNN'S)

The GNNs look into the propagation of dis-information in social networks, where the nodes represent the users and edges represent their interaction. It identifies clusters that demonstrate suspicious activities and finds traces of coordinated dis-information efforts by applying update rules of the node features, which could be formulated as follows:

$$\mathbf{h}_v^{(k)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \mathbf{W}^{(k)} \mathbf{h}_u^{(k-1)} + \mathbf{b}^{(k)}\right)$$

D. Multimodal Detection

One of the major advancements in enabling in-depth content analysis has been the integration of varied data formats such as text, images, and audio/video. This method is instrumental in uncovering different kinds of manipulations, for example, deepfakes, wrongly labeled images, and large-scale disinformation campaigns that are designed for particular platforms.

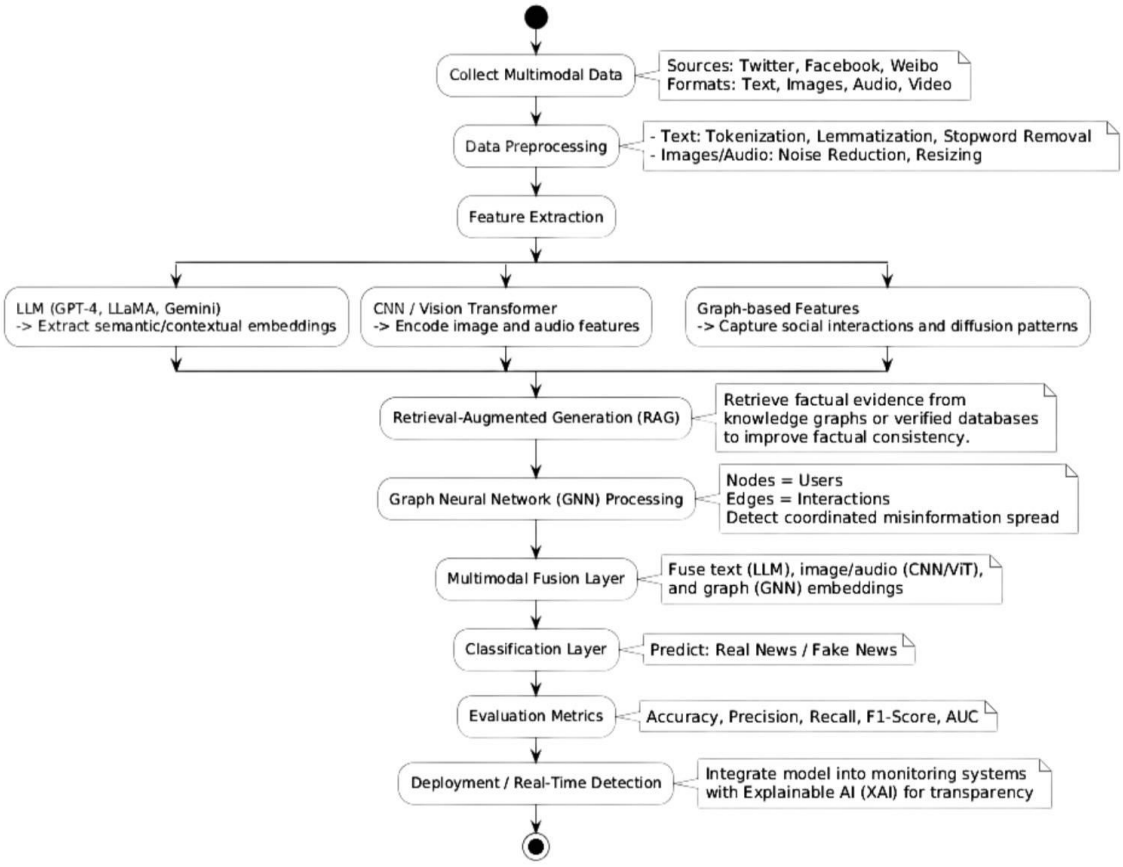


Figure. End-to-End Workflow of Fake News Detection Using LLM, RAG, and GNN Integration

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

V. METHODOLOGY

The methods employed in sorting out dangerous data generally require the use of an arranged approach that changes primary data from social situations into a type fit for categorizing. The complex operations, as per the guidelines of the background studies and the project requirements [1][2][3], include the subsequent steps:

1. **Data Acquisition:** Information has been gathered from a large number of online social networking sites, for example, Twitter, Facebook, and Weibo, and also from popular datasets like ISOT, FakeNewsNet, and PHEME. These collections can have a variety of content types, e.g., articles, posts, comments, and, in some instances, data regarding the spreading of the news from different networks.
2. **Data Preprocessing:** The unprocessed data from social media often include unnecessary things like hashtags, URLs, stopwords, and emojis; therefore, preprocessing is an important part. Standard operations include tokenization, lemmatization, removal of stopwords, and normalization. For multimodal datasets, preprocessing may also involve enhancing images and changing their sizes to make feature extraction easier.
3. **Feature Extraction:**
    - Traditional methods: Bag-of-Words (BoW), TF-IDF, and N-grams.
    - Deep learning methods: Word embeddings such as Word2Vec, GloVe, and contextual embeddings from transformer models like BERT.
    - Graph-based methods: Features derived from user–post interaction networks to capture diffusion patterns.
4. **Model Training and Detection:** The features that are extracted serve as the basis for the training of machine learning classifiers (for instance, Logistic Regression, SVM, Random Forest) as well as deep learning models (for example, CNNs, LSTMs, transformer architectures). Contemporary methods use a blend of Large Language Models (LLMs) like GPT-4 or LLaMA with Graph Neural Networks (GNNs) and Retrieval-Augmented Generation (RAG) to address the cases of multilingual, multimodal, and zero-shot situations.
5. **Evaluation:** Various metrics such as Accuracy, Precision, Recall, F1-score, and AUC are employed to evaluate the performance of the different models. To guarantee their strength, training, validation and test splits or cross-validation methods are used.
6. **Deployment and Real-Time Detection:** In the case of on-the-ground scenarios, models are integrated into surveillance systems that can handle live-stream data. In such workflows, the input is first preprocessed, then classified as genuine or fabricated, and in most cases, Explainable AI (XAI) techniques are used for providing better understanding and building the user's confidence.

This systematic pipeline ensures that fake news detection systems are changed to those accepting unstructured raw data to ones delivering real-time, explainable, and scalable classification suitable for social media environments.

VI. STRENGTHS AND LIMITATIONS

A. Strengths

- **High Accuracy:** Advanced models like Transformers and LLMs achieve near-human accuracy (95–99%) on benchmark datasets [6][7].
- **Automation at Scale:** AI makes it possible to have a large-scale, real-time social media platform monitoring, which is beyond the capability of manual moderation, by a human.
- **Context Awareness:** The use of deep learning and self-attention mechanisms allows the system to capture semantic relationships thus, it becomes more reliable in classification.
- **Multimodal Capabilities:** Integration of text, images, and metadata enhances detection performance in diverse content types.
- **Cross-Lingual Adaptability:** Multilingual embeddings and LLMs expand coverage to multiple languages, reducing language barriers.
- **Explainability with XAI:** Integration of explainable AI tools increases transparency and helps policymakers trust system outputs.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

B. Limitations

- **Dataset Bias:** Limited and imbalanced datasets often reduce model generalization to unseen or real-world data [2].
- **Computational Cost:** Transformer and LLM-based models demand heavy GPU/TPU resources for training and deployment.
- **Adversarial Manipulation:** The creators of fake news redesign their methods, thus, the detection systems become less effective against the newly unfolded misinformation.
- **Domain Dependency:** Research models, which are developed for a particular domain (e.g., COVID-19), normally, do not perform well in the case of different topics.
- **Delay in Real-Time Operation:** The performance of the classification can be slowed down by the use of a complicated model, thus, the real-time scalability is affected.
- **Black-Box Nature:** Even with XAI, deep learning models are still partially interpretable, which imposes a limitation to the trust that users place in them.

VII. DATASETS FOR FAKE NEWS DETECTION

Dataset	Content Type	Usage in Detection	Year	Remarks
FakeNewsNet	News articles + user interactions	Combines article content with social context for accurate detection	2018	Widely used benchmark dataset [1]
Twitter15/16	Tweets + retweets	Analyzes rumor spread patterns in real-time	2015/2016	Early benchmark for rumor detection [2]
PHEME	True/false/rumor threads	Classifies rumors and user stances during crisis events (e.g., protests, shootings)	2016	Popular for stance-based detection [3]
Weibo	Chinese rumors (text + images)	Enables multilingual and multimodal fake news detection	2017	Expands research beyond English datasets
Kaggle Fake News	English news articles (real/fake)	Common ML benchmarking dataset for binary classification	2017	Entry-level benchmark [4]
CoAID	COVID-related fake news	Focuses on health misinformation during the pandemic	2020	Domain-specific (health) [5]
COVID-HeRA	COVID health impact misinformation	Measures severity and harm of misleading health claims	2021	Specialized health misinformation dataset
MediaEval	Multimedia (text + visual)	Detects manipulated images/videos alongside misleading captions	2021	Supports multimodal detection
CheckThat-2022	Human + AI-generated claims	Targets ChatGPT/LLM-generated misinformation and fact verification	2022	Latest benchmark for AI-era misinformation
WELFake	News articles (balanced dataset)	Balanced dataset addressing imbalance issues in fake/real news classification	2020	Improved class balance [6]
LIAR	Short political statements	Labels claims on a 6-point scale (true → pants-on-fire)	2017	Focused on politics [7]
PolitiFact	Fact-checked news articles	Used for political fake news classification	2016	Cited widely in misinformation research

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 10, October 2025

|DOI: 10.15680/IJIRSET.2025.1410091|

VIII. FUTURE SCOPE

Future research in fake news detection is expected to move toward more intelligent, scalable, and explainable systems:

- **Multimodal Integration:** Future models will combine text, images, videos, and even audio to detect misinformation more reliably.
- **Cross-Lingual and Low-Resource Languages:** Expanding detection to regional and low-resource languages is crucial for global impact.
- **Real-Time Detection Pipelines:** Deployment-ready systems with lightweight architectures will enable large-scale monitoring with minimal latency.
- **Explainable AI (XAI):** Enhancing transparency and trust by making model predictions interpretable for users and policymakers.
- **Graph + LLM Synergy:** Combining Graph Neural Networks (GNNs) with Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) will improve credibility checks by analyzing both content and propagation.
- **Adversarial Robustness:** Future systems will defend against adversarial attacks where fake news creators manipulate text or images to evade detection.
- **Human–AI Collaboration:** Detection tools will increasingly support journalists, fact-checkers, and regulators rather than replacing them.

IX. CONCLUSION

The identification of fake news has developed from basic machine learning models to sophisticated LLMs, RAG, and graph-based methods that can almost match human accuracy. Although these innovations seem very promising, there are still issues such as dataset bias, multilingual detection, multimodal content, and real-time scalability. The next investigation in this area should be more committed to explainable AI, efficient architectures, and stronger benchmarks for the provision of a dependable and transparent fake news detection service on social media.

REFERENCES

[1] S. Kuntur, A. Wróblewska, M. Paprzycki, and M. Ganzha, “Under the Influence: A Survey of Large Language Models in Fake News Detection,” IEEE Transactions on Artificial Intelligence, vol. 6, no. 2, pp. 112–130, Feb. 2025.
[2] A. Bhardwaj and S. Kim, “Fake Social Media News and Distorted Campaign Detection Using Sentiment Analysis and Machine Learning,” Heliyon, vol. 10, no. 8, pp. 1–10, Aug. 2024.
[3] M. V. Sanida, T. Sanida, A. Sideris, M. Dossis, and M. Dasygenis, “Fake News Detection Approach Using Hybrid Deep Learning Framework,” in Proc. 9th SEEDA-CECNSM, 2024, pp. 1–6.
[4] A. Wang and T. Gu, “Exploring the Convergence of Generative AI and Fake News Detection: Technological Advancements and Challenges,” in Proc. IEEE AINIT Conf., 2025, pp. 55–62.
[5] R. Omowaiye, I. Ghafir, M. Lefoane, S. Kabir, A. Qureshi, and M. R. Daham, “Artificial Intelligence and Big Data Analytics for the Detection of Fake News on Social Media,” in Proc. ICECCME, Maldives, 2024, pp. 1–7.
[6] B. Thuraisingham and T. Thomas, “Social Media Governance and Fake News Detection Integrated with Artificial Intelligence Governance,” in Proc. IEEE Int. Conf. on Information Reuse and Integration (IRI), 2024, pp. 220–227.
[7] C.-O. Truică, R. M. Drăgușin, A. Dumitrașcu, and A.-F. Danciu, “DANES: Deep Neural Network Ensemble for Social and Textual Context-Aware Fake News Detection,” arXiv preprint arXiv:2302.01456, Feb. 2023.
[8] A. Amira, A. Derhab, S. Hadjar, M. Merazka, Md. G. R. Alam, and M. M. Hassan, “Detection and Analysis of Fake News Users’ Communities in Social Media,” IEEE Transactions on Computational Social Systems, vol. 11, no. 4, pp. 987–1001, Aug. 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details