Fake News Detection in Social Media

Sanket Mohan Kotkar
Computer Engineering Department, DYPCOE, Akurdi, Pune, Maharashtra, India -411044
kotkarsanket1606@gmail.com

Abstract— The rise of social media has reshaped information flow, but it has also accelerated the spread of fake news. Such misinformation impacts politics, healthcare, the economy, and social trust, where manual moderation alone is insufficient. Artificial intelligence has therefore become central to detection efforts. Initial research used traditional machine learning methods like Logistic Regression, Naïve Bayes, and SVM, which provided only shallow analysis. Later, deep learning techniques such as CNNs and LSTMs enhanced semantic and sequential understanding. Recently, transformer-based models and Large Language Models (LLMs), including BERT, RoBERTa, and GPT-4, have delivered state-of-the-art performance. This survey reviews detection methods, benchmark datasets, and major developments from 2016 to 2025. It further examines strengths, weaknesses, and ongoing challenges such as dataset bias, multilingual coverage, and scalability, while pointing to future directions like multimodal detection and explainable AI.

Keywords— Fake News Detection, Social Media, Artificial Intelligence, Machine Learning, Deep Learning, Transformer Models, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Graph Neural Networks (GNNs), Multimodal Detection, Explainable AI (XAI)

I. INTRODUCTION

In the current digital landscape, platforms such as Facebook, Twitter, WhatsApp, and YouTube have become key sources of news for millions across the globe. Their rapid dissemination, user-friendly interface, and interactive capabilities make them effective communication tools; however, they also facilitate the rampant spread of misinformation. "Fake news," which refers to intentionally deceptive or fabricated information presented as legitimate news, poses significant threats to democracy, public confidence, and social cohesion [1], [2]. Unlike conventional media that is governed by editorial controls and regulations, social media operates without centralized oversight, enabling malicious individuals to quickly disseminate false narratives on a large scale [3].

The impact of misinformation can be seen in various global and regional incidents including the 2016 U.S. elections, the COVID-19 pandemic, and the 2020 Delhi riots. These instances illustrate how erroneous information can influence political results, endanger lives, incite violence, and lead to economic turmoil [4]. Given the vast quantity and variety of online content generated daily, manual detection methods are impractical; thus advanced computational techniques are essential [5], [6]. This review explores the development of fake news detection methodologies, examines commonly utilized datasets, and charts technological advancements from traditional machine learning approaches to contemporary large language models (LLMs) [7], [8], while also indicating potential future research directions and real-world applications.

II. NEED FOR AUTOMATED DETECTION

- **High Volume of Data**: Social media platforms generate millions of posts every minute, making it impossible for human moderators to keep up.
- **Rapid Spread of Misinformation**: The velocity at which fake news circulates often exceeds the response rate of fact-checking organizations—particularly during elections or health emergencies.
- **Limitations of Human Moderation**: Manual detection processes are slow and

costly while being susceptible to biases or oversights.

- **Scalability**: Automated systems can efficiently monitor extensive networks in real time.
- **Multimodal and Multilingual Coverage**: AI models possess the capability to simultaneously analyze text, images, videos, and multiple languages.
- **Consistency and Reliability**: Automation minimizes human error by ensuring uniform decision-making across different contexts.

III. EVOLUTION OF FAKE NEWS DETECTION TECHNIQUES

The progression in fake news detection reflects a shift from basic machine learning classifiers toward sophisticated Large Language Models (LLMs), further enhanced by graph-based techniques and retrieval augmentation. Each evolutionary phase has systematically addressed the shortcomings of earlier approaches while progressively improving accuracy, robustness, and adaptability—particularly in tackling multilingual and multimodal misinformation challenges [1], [2].

A. Traditional Machine Learning (2016–2017)

In its early stages, fake news detection predominantly relied on classical machine learning algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machines (SVMs), Random Forests, and K-Nearest Neighbors (KNNs) [1]. These algorithms employed handcrafted features like Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and N-grams for text representation. While these models were interpretable and computationally efficient for basic detection tasks, they struggled to capture deeper semantic meanings and contextual relationships between words, which limited their effectiveness [2].

- **Techniques Involved:** Logistic Regression, Naïve Bayes, SVMs, Random Forests, KNNs; BoW; TF-IDF; N-grams.
- **Accuracy Achieved:** 70–85% (Datasets: ISOT, Kaggle Fake News).

B. Deep Learning Approaches (2018–2019)

To address inadequacies in feature representation found in traditional models, researchers began implementing deep learning strategies such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Bidirectional LSTMs (Bi-LSTMs) [2], [3]. These models employed embeddings like Word2Vec and GloVe to better capture semantic similarities between words. CNNs were particularly effective at extracting local textual features, while LSTMs excelled at modeling long-range dependencies within text sequences. This significantly reduced reliance on handcrafted features and improved performance across varied datasets [4].

- **Techniques Involved:** CNNs, LSTMs, Bi-LSTMs, Word2Vec, GloVe.
- **Accuracy Achieved:** 85–92% (Datasets: Twitter15, Twitter16, Weibo).

C. Hybrid Deep Learning Models (2020–2021)

As misinformation became increasingly intricate on social media platforms, hybrid models that integrated various neural architectures gained traction. Techniques such as CNN combined with LSTM or Bi-LSTM, alongside Gated Recurrent Units (GRUs), emerged together with attention-based networks [4]. CNN layers focused on capturing local word-level features, while sequential dependencies were effectively addressed by LSTMs. The incorporation of attention mechanisms further allowed these models to prioritize significant words or phrases, thereby enhancing detection capabilities across both short-form texts and longer documents. These hybrid systems consistently outperformed standalone models when classifying event-driven rumors or viral content [5].

- **Techniques Involved:** CNN+LSTM, Bi-LSTM+GRU, Attention Mechanism, FastText.
- **Accuracy Achieved:** 90–95% (Datasets: FakeNewsNet, PHEME, ISOT).

D. Transformer-Based Models (2022–2023)

The advent of transformer architectures represented a pivotal advancement in fake news detection methodologies. Pretrained transformer frameworks such as BERT, RoBERTa, XLNet, ALBERT, and DistilBERT have established themselves as foundational architectures for tackling tasks related to fake news identification [6]. By leveraging self-attention mechanisms, these transformers significantly outperform LSTMs in recognizing long-range dependencies and capturing nuanced contextual information within texts. Moreover, fine-tuning these architectures on domain-specific datasets has led to substantial improvements in accuracy while minimizing overfitting issues, thereby establishing transformers as cutting-edge solutions for misinformation classification [7].

- **Techniques Involved:** BERT, RoBERTa, XLNet, ALBERT, DistilBERT, VGG19 (multimodal).
- **Accuracy Achieved:** 92–97% (Datasets: CoAID, COVID-HeRA, FakeNewsNet, Weibo).

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

Recent advancements encompass integrating Large Language Models (LLMs) such as GPT-4, ChatGPT, LLaMA, and Gemini with Graph Neural Networks (GNNs) alongside Retrieval-Augmented Generation (RAG) [8]. The deployment of LLMs facilitates zero-shot, few-shot, and multilingual capabilities, whereas GNNs analyze patterns involving how misinformation propagates through social connections. Furthermore, RAG strengthens verification processes by simultaneously cross-referencing claims against reliable references in real time. Collectively, these systems achieve near-human-level accuracy while exhibiting transparency through Explainable AI (XAI)—marking them as leading-edge innovations in current research on combating fake news [9].

- **Techniques Involved:** GPT-4, ChatGPT, LLaMA, Gemini, Retrieval-Augmented Generation (RAG), Graph Neural Networks (GNNs), FakeBERT, HGCN.
- **Accuracy Achieved:** 97–99.2% (Datasets: CheckThat-2022, MediaEval, FakeNewsNet, Twitter15/16).

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 most recent innovations in fake news detection leverage **Large Language Models (LLMs)**, **Graph Neural Networks (GNNs)**, and **Retrieval-Augmented Generation (RAG)**. These approaches go beyond traditional classification by combining **content analysis, propagation modeling**, and **external fact verification**, collectively achieving **near human-level accuracy** on benchmark datasets, with performance ranging from **97% to 99.2%**.

A. Large Language Models(LLM's)

Models such as GPT-4, LLaMA, and Gemini leverage billions of parameters, enabling enhanced contextual understanding. This allows them to operate effectively in zero-shot and few-shot scenarios, as well as perform multilingual detection across diverse domains. In particular, these models excel at detecting sarcasm, implicit biases, and subtle cues indicative of misleading information. A key mechanism underlying these transformer-based models is self-attention, which can be formulated mathematically as:

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

B. Retrieval-Augmented Generation(RAG)

Retrieval-Augmented Generation (RAG) enhances the capabilities of LLMs by retrieving relevant evidence from trusted external sources, such as knowledge graphs or verified databases, prior to generating responses. This process significantly reduces hallucinations while improving factual consistency during assessments, which are typically modeled probabilistically as follows:

$$P(y|x) = \sum_z P(y|x,z)P(z|x)$$

where z is the retrieved evidence.

C. Graph Neural Networks(GNN'S)

GNN's analyze the mechanisms behind misinformation dissemination across social networks, where nodes represent users and edges represent interactions. This structure enables the identification of clusters of suspicious activity, highlighting their potential to uncover coordinated disinformation campaigns, based on node feature update rules:

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

D. Multimodal Detection

Integrating diverse data types—namely text, images, and audio/video—constitutes a significant breakthrough by enabling holistic content analysis. This approach facilitates the detection of manipulations, including deepfakes, mislabeled images, and coordinated disinformation campaigns that unfold across multiple platforms.

V. METHODOLOGY

The detection of fake news typically follows a structured workflow that converts raw social media data into meaningful insights for classification. The adapted workflow, based on prior research and project frameworks [1], [2], [3], consists of the following stages:

1. **Data Collection:** Datasets are obtained from platforms such as Twitter, Facebook, and Weibo, or from established repositories including ISOT, FakeNewsNet, and PHEME. These datasets may contain articles, posts, comments, and information about how news spreads across networks.
2. **Data Preprocessing:** Since raw social media data often includes noise like hashtags, URLs, stopwords, and emojis, preprocessing is essential. Common steps include tokenization, lemmatization, stopword removal, and normalization. For multimodal datasets, preprocessing may also involve cleaning images and adjusting their dimensions for feature extraction.
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:** Extracted features are used to train machine learning classifiers (e.g., Logistic Regression, SVM, Random Forest) and deep learning models (e.g., CNNs, LSTMs, transformer architectures). Modern approaches combine Large Language Models (LLMs) such as GPT-4 or LLaMA with Graph Neural Networks (GNNs) and Retrieval-Augmented Generation (RAG) for handling multilingual, multimodal, and zero-shot scenarios.

5. **Evaluation:** Models are assessed using metrics like Accuracy, Precision, Recall, F1-score, and AUC. To ensure robustness, training, validation, and test splits or cross-validation strategies are applied.

6. **Deployment and Real-Time Detection:** For real-world applications, models are deployed in monitoring systems that process incoming posts in real time. These pipelines apply preprocessing, classify content as real or fake, and often incorporate Explainable AI (XAI) methods to improve transparency and trust.

This systematic pipeline ensures that fake news detection systems evolve from handling unstructured raw data to 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 enables large-scale, real-time monitoring of social media platforms, far beyond manual moderation.

- **Context Awareness:** Deep learning and self-attention mechanisms capture semantic relationships, improving classification reliability.

- **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.

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:** Fake news creators adapt strategies, making detection systems vulnerable to evolving misinformation.

- **Domain Dependency:** Models trained on specific domains (e.g., COVID-19) often fail when applied to unrelated topics.

- **Latency in Real-Time Use:** Complex models can slow down classification, affecting real-time scalability.

- **Black-Box Behavior:** Despite XAI, deep learning models still lack full interpretability, limiting user trust.

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

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

Fake news detection has advanced from simple machine learning models to powerful LLMs, RAG, and graph-based methods, achieving near-human accuracy. While these technologies show great promise, challenges like dataset bias, multilingual detection, multimodal content, and real-time scalability remain. Future research should emphasize explainable AI, efficient architectures, and stronger benchmarks to ensure reliable and transparent fake news detection 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.