

Artificial Intelligence and Fake News Detection on Social Media - A Review of Techniques, Challenges, and Future Directions

Dr. Pawar Santosh Barku Prof. Ghangale Vivek Vinayak Wagh Ajinkya Annasaheb

Vidya Niketan Institute of Management and Computer Applications, Bota, Sangamner

Abstract: *The rapid growth of social media platforms has transformed the way people create, access, and share information. However, the same platforms have facilitated the widespread dissemination of fake news, misinformation, and disinformation. False information can influence public opinion, create social unrest, damage reputations, and undermine trust in news organizations. Traditional manual fact-checking methods are often time-consuming and difficult to scale to the enormous volume of online content. Artificial Intelligence (AI) offers promising approaches to identifying potentially misleading information through automated content analysis, pattern recognition, natural language processing, and social network analysis.*

Contextual language representations and the integration of textual, visual, and social information can improve detection capabilities, but important challenges remain, including limited multilingual resources, changing misinformation patterns, model explainability, dataset bias, adversarial manipulation, and reliable real-time detection. The review concludes that future systems should combine AI-based screening with evidence verification, transparent explanations, human oversight, and privacy preserving practices..

Keywords: Artificial Intelligence, Fake News Detection, Social Media, Machine Learning, Deep Learning, Natural Language, Processing, Transformers, Multimodal Learning, Misinformation

I. INTRODUCTION

Social media platforms have become important sources of news, public discussion, and real-time information. Users can publish and share information rapidly, often without the editorial review associated with traditional journalism. Although this openness provides substantial benefits, it also allows misleading or fabricated information to spread widely.

Fake news generally refers to false or misleading information presented in a news-like format. Related concepts include misinformation, which is false information shared without necessarily intending to deceive, and disinformation, which involves deliberate deception. These distinctions matter because the factual accuracy of a claim and the intention of the person sharing it are separate questions.

Artificial Intelligence provides techniques for analyzing large amounts of online information. Machine learning models identify statistical patterns in textual and behavioural features, while deep learning models learn more complex representations from data. Natural language processing supports semantic analysis, and graph-based approaches examine relationships among users, posts, and information sources. Multimodal systems combine text with images, video, audio, and social context.

1.1 Objectives of the Review

- To examine the role of AI in detecting fake news on social media.
- To review traditional machine learning, deep learning, transformer, and graph-based methods.



- To investigate multimodal detection and evidence-based verification techniques.
- To summarize important datasets and evaluation metrics.
- To identify the limitations of existing approaches.
- To propose future research directions for reliable and explainable detection systems.

II. REVIEW METHODOLOGY

This paper adopts a narrative literature-review approach. It synthesizes established research on fake news detection and incorporates survey papers addressing deep learning, multimodal systems, and large language models. Relevant studies were identified through scholarly search resources associated with IEEE Xplore, ACM Digital Library, Springer, and related academic publication sources. Search concepts included fake news detection, AI-based misinformation detection, deep learning for social media, transformer-based classification, multimodal fake news detection, and social network analysis.

III. BACKGROUND AND CONCEPTUAL FRAMEWORK

3.1 Fake News and Related Concepts

Fake news detection is commonly treated as a classification or verification problem, but several related tasks should be distinguished.

- Fake news classification: Predicts whether a news item belongs to a defined false or unreliable category.
- Rumour detection: Identifies claims circulating without confirmed verification.
- Claim verification: Compares a factual claim with supporting or contradicting evidence.
- Source credibility assessment: Evaluates the reliability of a publisher or information source.
- Manipulated media detection: Identifies altered, synthetic, or misleading visual and audio content.
- Stance detection: Determines whether a response supports, denies, questions, or comments on a claim.

3.2 Why Social Media Makes Detection Difficult

Social media creates a dynamic information environment in which content is copied, edited, translated, and shared. Popularity does not establish accuracy, and early engagement may reflect emotional appeal rather than credibility. Additional complications include automated accounts, coordinated campaigns, satire, impersonation, manipulated screenshots, and AI-generated text or media.

IV. ARTIFICIAL INTELLIGENCE TECHNIQUES FOR FAKE NEWS DETECTION

4.1 Traditional Machine Learning

Early automated systems commonly used manually designed textual and statistical features, including word frequency, TF-IDF, sentiment, punctuation, headline characteristics, writing style, and source-related metadata. Common algorithms include Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, and gradient-boosting methods.

The general workflow consists of collecting labelled examples, cleaning and preprocessing text, extracting features, training a classifier, and evaluating predictions on held-out data. Traditional machine learning offers relatively low computational cost and useful baselines, but may struggle with context, irony, complex reasoning, and claims that are misleading despite appearing linguistically credible.

4.2 Deep Learning

Deep learning models automatically learn representations from large datasets and can capture more complex patterns than many feature-engineering approaches. Convolutional Neural Networks (CNNs) learn local patterns in text and images.



Recurrent Neural Networks (RNNs) process sequences, while Long Short-Term Memory (LSTM) networks are designed to retain relevant information over longer sequences. Attention mechanisms help models emphasize relevant parts of a post or conversation.

Deep learning can reduce dependence on manually selected features, but effectiveness depends on training-data quality, model design, and evaluation methodology. Models can learn accidental correlations, such as publisher names or topic-specific vocabulary, instead of genuine indicators of factual accuracy.

4.3 Transformer-Based Models

Transformer architectures have substantially influenced NLP by enabling contextual representations of words and long-range relationships within text. Models such as BERT and RoBERTa can be fine-tuned on labelled examples. Unlike simple wordfrequency methods, contextual models can represent a word differently depending on its surrounding sentence.

Potential advantages include improved semantic representation and transfer learning. Nevertheless, a transformer does not automatically verify facts. A model trained on historical labels may confidently misclassify new claims, especially when evidence is incomplete or the subject is unfamiliar. Transformer-based detection is best understood as identifying potentially suspicious content rather than replacing fact-checking.

4.4 Graph Neural Networks and Social Context

News spreads through networks of users, posts, replies, reposts, and sources. Graph-based models represent these relationships as nodes and edges. A graph may contain users and interactions, news articles and associated posts, sharing relationships, replies expressing agreement or disagreement, and connections between claims and external evidence.

Graph Neural Networks aggregate information from connected nodes to learn representations of content and propagation structure. These methods may detect coordinated dissemination that text-only models miss. However, legitimate breaking news can spread rapidly, and coordinated false content may resemble ordinary discussion. Social context should be treated as supporting evidence, not definitive proof.

4.5 Multimodal Fake News Detection

Online misinformation often combines written claims with photographs, screenshots, videos, audio, and graphics. A text-only classifier may miss a contradiction between a caption and accompanying media. Multimodal systems integrate text, images, video and audio, social context, and external knowledge.

Feature fusion can occur at the input level, representation level, or decision level. For example, a system may encode a headline with a language model, encode an image with a vision model, and combine both representations to evaluate whether the image supports the headline. These approaches are promising, but introduce data-alignment problems, computational costs, and explanation challenges [3], [4].

4.6 Large Language Models and Retrieval-Augmented Verification

Large Language Models (LLMs) can assist with interpreting claims, extracting entities, identifying factual assertions, and summarizing evidence. Retrieval-augmented systems search trusted documents, official sources, and fact-checking records, then compare retrieved evidence with a claim.

- Extract the principal factual claim from a social media post.
- Identify entities, dates, places, and events.
- Retrieve relevant evidence from trusted sources.
- Assess whether evidence supports, contradicts, or fails to resolve the claim.
- Generate an explanation with links to supporting evidence.
- Refer uncertain or high-impact cases to a human reviewer.



V. LITERATURE REVIEW

5.1 Deep Learning for Fake News Detection

Hu et al. reviewed deep learning-based fake news detection, including supervised, weakly supervised, and unsupervised approaches [1]. Their review examines how content, social context, and external knowledge contribute to detection. Deep learning can extract complex representations, but effectiveness depends on training data, label quality, and realistic evaluation.

5.2 Multimodal Detection

Comito et al. examined multimodal fake news detection and highlighted the importance of combining information from different media types [3]. Hangloo and Arora reviewed methods, datasets, and future perspectives for combating multimodal fake news [4]. These studies highlight multimodal fusion, cross-domain generalization, and limited training data as important research concerns.

5.3 Social Context and Dataset Development

Shu et al. introduced FakeNewsNet, a data repository designed to support research using news content, social context, and spatiotemporal information [6]. Such datasets help researchers investigate how the properties of a news item and its dissemination environment relate to classification. However, labels can reflect the decisions of specific fact-checking organizations, collection periods, or topic distributions.

5.4 Systematic Reviews and Emerging Models

Mostafa et al. reviewed unimodal and multimodal deep learning methods, feature extraction, evaluation measures, and challenges posed by large language models [2]. Elgammal and Alhajj's 2026 review examines deep learning for textual misinformation detection and emerging LLM applications [5]. The literature indicates growing interest in contextual language models, transfer learning, multimodal analysis, and evidence-aware methods.

5.5 Summary of the Literature

- From handcrafted features to learned representations: Deep learning and transformers provide more flexible representations of language.
 - From isolated content to contextual analysis: Social networks, user interactions, and propagation patterns provide additional signals.
 - From classification to evidence-based verification: Retrieval, multimodal reasoning, and human review can help distinguish a suspicious prediction from a substantiated factual judgment.
- These developments are complementary. The appropriate approach depends on the task, available evidence, computational resources, and the consequences of an incorrect prediction.

VI. COMPARISON OF AI TECHNIQUES

Technique	Main input	Advantages	Limitations
Traditional ML	Text features and metadata	Efficient; useful baseline	Limited contextual understanding
CNN	Text or images	Learns local patterns	May miss broader context
LSTM/RNN	Text sequences	Models sequential information	Training and long-range limitations
Transformers	Contextual text	Strong semantic representations	Computational cost; does not inherently verify facts
GNNs	User and content networks	Models propagation and relationships	Sensitive to network quality and platform changes
Multimodal AI	Text, images, video, other signals	Combines complementary evidence	Data alignment and fusion complexity



LLMs with retrieval	Claims and evidence	external evidence	Can explain evidence and support verification	Hallucinations, retrieval errors, and latency
---------------------	---------------------	-------------------	---	---

No single method is universally best. Results depend on the dataset, task definition, class distribution, temporal setting, and evaluation protocol. Accuracy figures from different studies should not be compared directly unless experimental conditions are sufficiently similar.

VII. DATASETS USED IN FAKE NEWS DETECTION

Dataset	Main characteristics	Research application
LIAR	Short claims with fact-checking labels	Claim classification
FakeNewsNet	News content, social context, temporal information	Content and propagation analysis
PHEME	Rumours and related social-media conversations	Rumour verification and conversation analysis
Fakeddit	Multimodal social-media posts	Text-image analysis
ISOT Fake News Dataset	News articles labelled real or fake	Text classification baselines

The datasets differ in collection methods, label definitions, time periods, and content formats. Researchers should examine licensing, availability, platform restrictions, and label provenance. They should also check for duplicate articles, overlapping sources, class imbalance, and temporal leakage. Publisher-separated, time-separated, and cross-domain evaluations can provide a more realistic assessment of generalization.

VIII. EVALUATION METRICS

8.1 Accuracy

Accuracy measures the proportion of correctly classified examples:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

8.2 Precision

Precision measures the proportion of predicted fake-news cases that are actually labelled fake: $\text{Precision} = TP / (TP + FP)$. High precision can reduce false accusations against legitimate content.

8.3 Recall

Recall measures the proportion of labelled fake-news cases correctly identified:

$\text{Recall} = TP / (TP + FN)$. High recall can reduce the number of false items that escape detection.

8.4 F1-Score

The F1-score is the harmonic mean of precision and recall:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

8.5 Additional Evaluation Measures

- ROC-AUC: Ranking performance across classification thresholds.
- Precision-Recall AUC: Useful when positive examples are relatively rare.
- Calibration: Whether predicted confidence corresponds to observed correctness.
- Detection latency: Time required to identify potentially misleading content



IX. MAJOR CHALLENGES

9.1 Data Quality and Label Reliability

Fake news datasets often contain labels assigned by fact-checking organizations, researchers, or platform moderators. Standards can differ, and some claims remain unresolved when first published. A possible direction is to record the source and date of each label, preserve uncertainty categories, and periodically review outdated annotations.

9.2 Multilingual and Low-Resource Languages

Many resources and pretrained models have stronger coverage for high-resource languages. Regional languages, code-switching, transliteration, and local expressions create additional challenges. Future work should develop language-specific datasets and evaluation protocols that include regional varieties and mixed-language posts.

9.3 Changing Misinformation Patterns

Claims, writing styles, topics, and media-generation techniques evolve. Models trained on historical data may fail when the distribution changes. Time-based evaluation, performance monitoring, and carefully reviewed retraining can help address this problem.

9.4 Explainability

A classification label without an explanation provides limited value. Fluent explanations can also sound convincing while relying on incorrect evidence. Systems should link decisions to specific claims and verifiable sources, distinguish evidence from inference, and communicate unresolved questions.

9.5 Adversarial Attacks and AI-Generated Content

People may deliberately change wording, manipulate images, use synthetic media, or create coordinated accounts to evade automated systems. Combining content analysis with source verification, media forensics, robust training, and independent evidence retrieval may improve resilience.

X. PROPOSED FRAMEWORK FOR AI-BASED FAKE NEWS DETECTION

This paper proposes a conceptual framework combining machine learning, multimodal analysis, external evidence, and human oversight.

Stage 1: Data collection. Collect public posts, news articles, associated media, and relevant metadata in accordance with privacy rules and platform terms.

Stage 2: Preprocessing. Remove duplicates, normalize text, identify languages, extract claims, and prepare images or videos.

Stage 3: Feature extraction. Use a transformer for textual representations, a vision model for images, and graph-based features for relevant propagation patterns.

Stage 4: Initial classification. Use a lightweight classifier to estimate whether content requires additional review; treat this as a screening signal, not proof.

10.1 Conceptual Workflow

Raw social media content → preprocessing and claim extraction → AI screening → multimodal and social-context analysis → external evidence retrieval → evidence assessment → human review when required → transparent result.

10.2 Expected Benefits

- Improve identification of potentially misleading content.
- Combine textual and visual signals with relevant context.
- Reduce unsupported conclusions by grounding assessments in external evidence.



- Communicate uncertainty rather than forcing every claim into a binary category.

XI. FUTURE RESEARCH DIRECTIONS

- 11.1 Explainable and evidence-grounded AI: Identify the claims under assessment, cite supporting or contradicting evidence, and distinguish verified findings from model-generated interpretations.
- 11.2 Multilingual misinformation detection: Prioritize underrepresented languages, regional dialects, transliterated text, and code-switching, with native-language annotation and culturally appropriate resources.
- 11.3 Multimodal and cross-platform learning: Analyze text, images, audio, and video together while accounting for platform and media-format differences.
- 11.4 Detection of synthetic content: Study how AI-generated text, audio, and video affect misinformation detection. Provenance and forensic methods can complement other evidence, but no single detector identifies all synthetic content reliably.
- 11.5 Early detection: Assess claims before widespread circulation while accounting for limited evidence during breaking events.

XII. DISCUSSION

The reviewed research shows that AI offers complementary ways to address fake news detection. Traditional machine learning provides efficient baselines, deep learning learns complex representations, transformers support contextual text analysis, and graph models capture relationships within social networks. Multimodal systems examine inconsistencies between text and media, while retrieval-augmented LLM systems can assist with evidence-based claim assessment. A crucial distinction exists between detecting suspicious content and determining whether a claim is factually false. A model may learn correlations associated with previously labelled fake news without independently verifying a claim. High engagement, emotionally charged wording, or an unfamiliar source does not establish falsity.

XIII. CONCLUSION

Artificial Intelligence has become an important research area for detecting fake news on social media. Machine learning and deep learning can analyze large volumes of content, transformer architectures improve contextual language representation, graph-based methods examine dissemination patterns, and multimodal models combine text with images, audio, video, and social information. Retrieval-augmented language models offer further opportunities to support claim verification through external evidence.

Despite these advances, reliable detection remains difficult because misinformation evolves, datasets contain biases, factual evidence may be incomplete, and models can produce incorrect or misleading predictions. Multilingual coverage, explainability, privacy, real-time performance, and resistance to adversarial manipulation remain significant challenges.

REFERENCES

- [1] L. Hu et al., "Deep learning for fake news detection: A comprehensive survey," *AI Open*, vol. 3, pp. 133–155, 2022. doi: 10.1016/j.aiopen.2022.09.001.
- [2] M. Mostafa, A. S. Almogren, M. Al-Qurishi, and M. Alrubaian, "Modality Deep-learning Frameworks for Fake News Detection on Social Networks: A Systematic Literature Review," *ACM Computing Surveys*, vol. 57, no. 3, article 77, 2024. doi: 10.1145/3700748.
- [3] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: a survey of deeplearning techniques," *Social Network Analysis and Mining*, vol. 13, article 101, 2023. doi: 10.1007/s13278-023-01104-w.
- [4] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: methods, datasets, and futureperspective," *Multimedia Systems*, 2022. doi: 10.1007/s00530-022-00966-y.



- [5] Z. Elgammal and R. Alhaji, "Evolution of Deep Learning Models for Misinformation Detection in Social Media Textual Data: Background, Architectures, Datasets, and Emerging LLM Applications," *Social Science Computer Review*, published online March 26, 2026. doi: 10.1177/08944393261432625.
- [6] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020. doi: 10.1089/big.2020.0062.

