International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Fake News Detection Using Machine Learning: A Systematic Review

Gopal Sharma[1], Varsha Ahiwar[2], Abhishek Mishra[3], Harshit Singh[4], Avinash Rajak[5],

Rajneesh Shrivastava[6]

[1] Department of Computer Science AKS University, Satna, M.P. India
[2] Department of Computer Science AKS University, Satna, M.P. India
[3] Department of Computer Science AKS University, Satna, M.P. India
[4] Department of Computer Science AKS University, Satna, M.P. India
[5] Department of Computer Science AKS University, Satna, M.P. India
[6] Department of Computer Science AKS University, Satna, M.P. India
Corresponding Author Email: rajsp.shrivastava@gmail.com

ORCID: https://orcid.org/[0009-0006-5294-7054

How to Cite this Article:

Shrivastava, R., Rajak, A., Singh, H., Mishra, A., Ahiwar, V. & Sharma, G. (2026). Fake News Detection Using Machine Learning: A Systematic Review. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.167

https://doi.org/10.55041/ijcope.v2i5.167

Abstract—

The rapid proliferation of misinformation and fabricated content across digital platforms constitutes one of the most pressing challenges of the information age. Fake news not only distorts public discourse but also influences elections, financial markets, and public health outcomes. This review paper presents a comprehensive survey of machine learning (ML) and deep learning (DL) approaches proposed for automated fake news detection. We systematically examine techniques spanning classical models — Naïve Bayes, Support Vector Machines, Logistic Regression, and Random Forest — through to contemporary transformer-based architectures such as BERT, RoBERTa, and GPT variants. Additionally, we review graph-based propagation methods, multi-modal fusion approaches, and cross-lingual detection strategies. Key benchmark datasets including LIAR, FakeNewsNet, BuzzFeed, and ISOT are described alongside standard evaluation metrics. We discuss prevailing challenges such as adversarial attacks, class imbalance, domain shift, and the scarcity of labelled data, and outline promising future research directions including explainability, federated learning, and real-time detection pipelines.

Keywords— Fake News Detection, Machine Learning, Deep Learning, Natural Language Processing, BERT, Misinformation, Social Media, Text Classification, Transformer Models

I. INTRODUCTION

The digital revolution has fundamentally transformed how information is created, shared, and consumed. Social networking platforms such as Facebook, Twitter, and WhatsApp now serve as primary news sources for billions of users worldwide. While these platforms democratize information access, they simultaneously enable the unchecked dissemination of misleading, fabricated, or deliberately deceptive content commonly referred to as 'fake news'. A landmark MIT study demonstrated that false information spreads approximately six times faster on Twitter than verified news, underscoring the urgency of automated detection solutions. [1]

Fake news encompasses a wide spectrum of deceptive content: satire and parody mistaken for facts, misleading framing of genuine events, imposter content using the branding of legitimate outlets, fabricated content with no factual basis, and manipulated media including deepfakes. The consequences are far-reaching — from swaying electoral outcomes and inciting social unrest to endangering lives through health misinformation, as witnessed during the COVID-19 infodemic.[2]

Manual fact-checking by human experts remains the gold standard but is fundamentally unscalable given the volume of content generated online. Consequently, the research community has invested heavily in automated detection systems leveraging natural language processing (NLP) and machine learning (ML). This paper provides a systematic review of the state of the art, covering dataset construction, feature engineering, classical ML classifiers, deep learning architectures, and emerging multi-modal and graph-based approaches. [3]

A. OBJECTIVE:

The primary objectives of this survey are:
(i) to catalogue major public benchmark datasets and their characteristics;
(ii) to systematically review ML and DL methods proposed for fake news detection;
(iii) to compare reported performance across models and datasets;

(iv) to identify open challenges

II. LITERATURE REVIEW

Research on automated fake news detection gained momentum around 2016 following high-profile misinformation events. Scholars have approached the problem from multiple angles, broadly categorized as propagation-based, content-based, source-credibility based, and hybrid approaches. [5]

A. Classical Machine Learning Approaches

Early works relied on hand-crafted linguistic and stylometric features fed into traditional classifiers. Castillo et al. (2011) pioneered credibility assessment of Twitter posts using decision trees, achieving 86% accuracy. Potthast et al. (2018) employed Random Forest on hyperpartisan news corpora and demonstrated that writing style alone carries strong discriminative signal. Ahmed et al. (2018) compared TF-IDF representations with Naive Bayes, Logistic Regression, Linear SVM, and Random Forest on the LIAR dataset; Linear SVM achieved the highest accuracy of 92% on binary classification tasks. Conroy et al. (2015) conducted an early survey highlighting linguistic cue-based detection and the need for network analysis features. [6]

B. Transformer-Based Approaches

Pre-trained transformer language models have set new benchmarks across NLP tasks and fake news detection is no exception. Kula et al. (2020) fine-tuned BERT on multiple fake news datasets, achieving 98.5% accuracy on binary ISOT classification. Zhou et al. (2020) proposed SAFE, combining visual and textual encoders with cross-modal attention to detect multi-modal fake news. Zellers et al. (2019) trained Grover — a model that both generates and detects neural fake news — demonstrating that the best detector of machine-generated text is often a similar generative model. Popat et al. (2018) proposed DeClarE, using attention-based neural networks over web sources for claim verification.[7]

III. METHODOLOGY

This section describes the datasets, feature extraction strategies, and model architectures employed in the fake news detection.

A. Datasets

Table 1 summarizes the most widely used benchmark datasets. The LIAR dataset (Wang, 2017) consists of 12,836 short statements labelled across six veracity classes with rich metadata. FakeNewsNet (Shu et al.,2018) aggregates news from PolitiFact and GossipCop, providing both content and social context. The ISOT dataset contains approximately 44,000 news articles from Reuters (real) and fabricated sources (fake). BuzzFeed News and FakeNewsCorpus provide large scale corpora for training deep models.

B. Feature Extraction

Linguistic Features: Bag-of-Words (BoW) and TF-IDF representations capture lexical patterns. N-gram models (unigrams, bigrams, trigrams) encode local context. Psycholinguistic features from LIWC (Linguistic Inquiry and Word Count) quantify emotional tone, cognitive processes, and social references. Stylometric features include sentence length distributions, punctuation density, and readability scores (Flesch-Kincaid, Gunning Fog).

Semantic Features: Pre-trained embeddings (Word2Vec, GloVe, FastText) map tokens into dense continuous vector spaces encoding semantic similarity. Sentence-level embeddings from Universal Sentence Encoder (USE) and InferSent capture discourse-level meaning. Transformer contextual embeddings (BERT,RoBERTa) provide the richest representations, encoding bidirectional context.

Social and Metadata Features: User credibility scores, account age, follower counts, verified status, posting frequency, and engagement patterns (likes, shares, comments) constitute the social feature space. Propagation tree topology —depth, breadth, retweet speed — provides temporal spreading signals.

C. Model Architectures

The following pipeline represents the general framework evaluated in this review:

1. Pre-processing: Tokenization, stop-word removal, lemmatization, URL and mention stripping, and case normalization.

2. Feature Representation: TF-IDF /embedding layer / pre-trained transformer encoder, optionally augmented with metadata and social features.

Dataset	Size	Labels	Domain	Features
LIAR	12,836	6-class	Politics	Text, Speaker, Context
FakeNewsNet	~23,000	Binary	Politics/Gossip	Text, Image, Social
ISOT	~44,000	Binary	General News	Text Only
BuzzFeed	~2,300	4-class	Politics	Text, Social
FEVER	~185,000	3-class	Wikipedia	Claims, Evidence
FakeNewsCorpus	~9.4 M	Binary	Mixed	Text Only

Table 1: Summary of benchmark datasets for fake news detection.

IV. RESULTS AND DISCUSSION

Table 2 presents the comparative performance of representative ML and DL models across key benchmarks, consolidating results reported in original papers and reproducibility studies.

A. Challenges and Open Problems

Adversarial Robustness: Adversarial text attacks (TextFooler, BERT-Attack) can reduce BERT-based detector accuracy by 20–40%, exposing critical vulnerabilities in production systems.

Dataset Bias and Domain Shift: Models trained on political news transfer poorly to health misinformation or satirical content.

Explainability: Black-box deep learning predictions are insufficient for deployment in journalistic and policy contexts. LIME, SHAP, and attention visualization methods have been applied, but coherent user-facing explanations remain an open challenge.

Low-Resource and Multilingual Settings: Most high-quality labelled datasets are English-centric. Cross-lingual transfer via multilingual transformers consistently underperforms monolingual fine-tuned models by 4–8%.

Temporal Concept Drift: Misinformation narratives evolve rapidly. Static models become stale within months, necessitating online learning or periodic retraining with fresh annotated data.

Model	Dataset	Accuracy (%)	F1-Score (%)	AUC-ROC
Naive Bayes + TF-IDF	LIAR	58.4	56.1	0.71
Linear SVM + TF-IDF	ISOT	92.3	91.8	0.97
Random Forest	FakeNewsNet	81.7	80.9	0.89
XGBoost + Linguistic Feat.	LIAR	63.2	61.5	0.74
BiLSTM + GloVe	LIAR	72.8	71.3	0.82
CNN + Word2Vec	FakeNewsNet	87.4	86.9	0.93
BERT (fine-tuned)	ISOT	98.5	98.4	0.99
RoBERTa (fine-tuned)	LIAR	74.1	73.6	0.84
Bi-GCN (graph-based)	FakeNewsNet	90.1	89.5	0.95
XLM-RoBERTa (cross-lingual)	Multi-lang.	79.3	78.7	0.88

Table 2: Performance comparison of representative models on benchmark datasets.

V. CONCLUSION

This review has provided a comprehensive survey of machine learning and deep learning approaches to fake news detection. We traced the evolution of the field from hand-crafted feature-based classifiers to pre-trained transformer models and graph neural networks, cataloguing key datasets, methodological advances, and empirical benchmarks. The evidence overwhelmingly indicates that transformer-based models, particularly fine-tuned BERT and RoBERTa variants, However, real-world deployment introduces a raft of challenges — adversarial robustness, domain shift, temporal drift, multilingual coverage, and the critical remain largely unsolved. Future research should priorities:
(i) robust multimodal and cross-lingual detection frameworks;
(ii) explainable AI (XAI) techniques producing journalist-friendly evidence summaries;
(iii)federated and privacy-preserving learning to enable collaborative model training without centralizing sensitive user data;
(iv) continual learning to cope with temporal concept drift; and
(v) end-to-end real-time detection pipelines integrated into social media content moderation workflows.

As generative AI models become increasingly capable of producing convincing synthetic text and images, the arms race between fake news generators multi-disciplinary approach combining advances in ML with media literacy, regulatory frameworks, and platform governance will be essential to effectively combat the global misinformation epidemic.

ACKNOWLEDGMENT

I want to thank everyone who helped and advised me during this study effort from the bottom of my heart. I am really grateful to my supervisor for their helpful advice, support, and encouragement. I also want to thank my school for giving me the tools and resources I need. I want to thank my coworkers and friends for their helpful ideas and support. I appreciate my family for always pushing me and giving me moral support. Their support was very important to the success of this research. Finally, I want to thank everyone who helped with this work, either directly or indirectly

REFERENCES

[1] Ahmed,H., Traore,I., & Saad, S. (2018). Detecting opinion spam and fake news using n-gram analysis and semantic similarity. IEEE Transactions on Information Forensics and Security, 14(2), 439–452.

[2] Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. AAAI, 34(1), 549–556.

[3] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. WWW '11, 675–684.

[4] Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. ASIST Annual Meeting, 52(1), 1–4.

[5] Kula, S., Cholewa, M., & Foszczy■ski, P. (2020). Using the BERT language model for fake news detection. Journal of Information and Telecommunication, 5(2), 215–229.

[6] Meel, P., & Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. Expert
Systems with Applications, 153, 112986.

[7] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social

[8] Nakamura, K., Levy, S., & Wang, W. Y. (2020). r/Fakeddit: A new multimodal benchmark dataset for

[9] Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. EMNLP 2018.