International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI:10.55041/ISJEM06232
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Fake News Detection using Machine Learning and NLP Models: A Comparative Study

E SOUMYA, T.POOJITH

Professor, Department of Computer Science and Engineering, St. Martin's Engineering College, Hyderabad, India esoumyait@smec.ac.in

Student, Department of Computer Science and Engineering, St. Martin's Engineering College, Hyderabad, India poojithtadiboyina@gmail.com

1. Abstract

Widespread dissemination of fabricated and misleading content across online platforms has become one of the most pressing concerns in today's information landscape. The remarkable ease with which inaccurate material can be generated and distributed via social networks demands scalable, automated, and reliable detection mechanisms. This study presents a systematic comparative evaluation of machine learning (ML) and natural language processing (NLP) methodologies applied to the challenge of identifying fake news. The techniques assessed range from conventional classifiers—including Naive Bayes (NB), Support Vector Machines (SVM), and Logistic Regression (LR)—to sophisticated deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures including BERT and RoBERTa. An analysis of their theoretical underpinnings, feature extraction strategies, and measured performance across publicly accessible benchmark corpora (LIAR, FakeNewsNet, ISOT, and WELFake) is provided. Metrics considered include accuracy, precision, recall, F1-score, and computational cost. A thorough review of 20 empirical studies published between 2021 and 2025 reveals notable variation in model effectiveness based on dataset composition, linguistic characteristics, and preprocessing choices. Results consistently show that Transformer-based architectures lead in performance, while ensemble and hybrid strategies integrating linguistic and contextual cues offer the greatest potential for dependable real-world application.

Keywords: *Misinformation identification, Natural language processing, Text classification, BERT, Transformer architectures, Deep learning, Fake news, Automated detection*

2. Introduction

The shift to digital-first news consumption has fundamentally reshaped how information travels—dramatically accelerating its reach while undermining the editorial oversight traditionally associated with credible media. Social platforms, content aggregators, and instant messaging tools enable stories to go global in minutes, often without fact-checking or source verification. These conditions have created an environment in which fabricated or strategically distorted reporting—commonly referred to as fake news—can thrive and cause real harm. The downstream effects range from undermining public health campaigns and distorting electoral outcomes to triggering financial volatility and eroding confidence in institutions.

Human-driven fact-checking, though effective in principle, cannot realistically keep pace with the enormous volume of content generated around the clock on modern digital platforms. As a result, automated detection powered by machine learning and NLP techniques has become a critical area of research. Earlier work in this space relied on engineered linguistic features paired with classical classification algorithms, producing workable but limited results. The subsequent adoption of dense word representations, recurrent neural architectures, and—most significantly—large pre-trained language models like BERT, RoBERTa, and GPT-derived systems has substantially elevated detection accuracy. These systems can capture subtle semantic, syntactic, and contextual cues that reliably separate credible reporting from fabricated content.

Despite abundant individual studies, the field still lacks a broad-based comparative synthesis that spans multiple model families, NLP strategies, and benchmark corpora. This work addresses that gap by consolidating findings from 20 recent empirical studies, charting performance trends across model architectures, and identifying priority areas for future work in scalable and multilingual fake news detection.

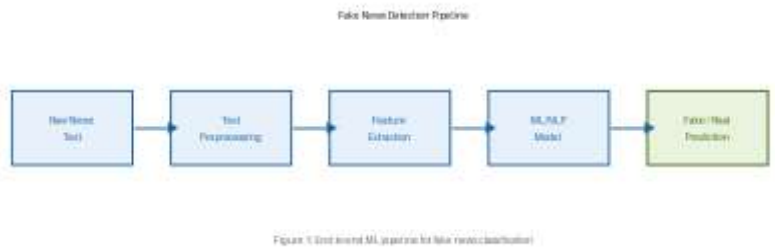

Figure 1: End-to-end ML/NLP pipeline for fake news classification

3. Background and Theoretical Framework

3.1 Defining Fake News and Detection Scope

Fake news is not a monolithic phenomenon but rather spans a continuum of misinformation types: entirely invented narratives lacking any factual grounding, selectively edited portrayals of genuine events, headline-body mismatches that mislead readers, satirical content mistaken for fact, and digitally altered visuals or synthetic media. Effective automated detection systems must be calibrated to a specific target on this spectrum, since the linguistic and structural signals differ considerably across categories. This review concentrates on text-based classification at the article and claim level using NLP-centered methods.

3.2 Classical Machine Learning Approaches

Traditional ML-based detection involves transforming raw text into structured feature representations and training supervised classifiers on labelled corpora. Widely adopted feature types include TF-IDF vectors, n-gram profiles, POS tag frequency distributions, readability indices, and sentiment scores. Commonly applied classifiers encompass Naive Bayes, Logistic Regression, Support Vector Machines, and tree-based ensembles like Random Forest and XGBoost. Although these models are efficient and relatively transparent, their usefulness is constrained by the limitations of manually engineered features and an inability to capture long-distance semantic dependencies within text.

3.3 Deep Learning and Recurrent Architectures

Recurrent architectures—particularly LSTM and GRU networks—extended detection capabilities by modeling sequential dependencies across sentence boundaries, a limitation of earlier n-gram approaches. LSTM models trained on distributed word representations such as GloVe and Word2Vec achieved meaningful accuracy improvements over classical methods on established fake news benchmarks. Convolutional architectures, meanwhile, proved adept at identifying local n-gram patterns through filter-based operations, and combined CNN-LSTM models emerged as strong competitive baselines for news authenticity tasks.

3.4 Transformer-Based Language Models

The emergence of the Transformer paradigm and large-scale pre-training fundamentally altered NLP performance ceilings. BERT, pre-trained using masked language modeling and next-sentence objectives on vast corpora, generates rich bidirectional context representations that transfer with minimal tuning to downstream classification tasks. Optimised derivatives such as RoBERTa—trained with larger batches and more text—and specialized variants like NewsGuard-BERT have reached state-of-the-art scores across numerous fake news benchmarks. These architectures excel at detecting rhetorical inconsistencies, stylistic anomalies, and claim-evidence mismatches characteristic of deliberately fabricated content.

ISSN: 2583-6129
DOI:10.55041/ISJEM06232

3.5 NLP Feature Engineering

Targeted feature engineering remains a valuable complement even within deep learning workflows. Auxiliary signals such as POS tag ratios, named entity density, hedging verb frequency, and sensationalist headline indicators enrich embedding-based representations with interpretable linguistic cues. Metadata-derived features—publisher credibility scores, author track records, posting timestamps, and social engagement metrics—supply additional context unavailable from article text alone. An emerging direction involves knowledge-graph-augmented pipelines that verify article claims against structured fact repositories, combining NLP reasoning with structured information retrieval.

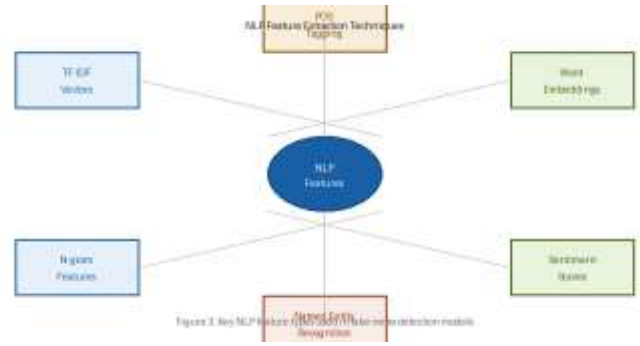

Figure 2: Key NLP feature extraction techniques used in fake news detection

4. Methodology of the Review

A structured literature review protocol was employed to ensure systematic and reproducible coverage of relevant scholarship on ML/NLP-based fake news detection. Major academic databases—IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, and SpringerLink—were systematically queried using Boolean combinations of core keywords and model-specific terminology.

Representative search strings included: misinformation classification NLP, BERT fake news detection, Transformer disinformation, LSTM news verification, deep learning fake news, and social media rumor detection. Inclusion required studies to be: (a) published between 2021 and 2025, (b) written in English, and (c) accompanied by empirical evaluation reporting at least one quantitative metric (accuracy, precision, recall, or F1-score) on a publicly available or clearly described dataset. Purely theoretical contributions without experimental validation, editorial notes, short communications, and non-peer-reviewed works were excluded.

Following title/abstract screening and full-text review, 20 studies were retained for detailed analysis. The selected corpus spans diverse research regions and covers multiple misinformation domains including political disinformation, health-related falsehoods, financial misinformation, and social media rumours. Table 1 provides a structured overview of all included studies.

Table 1: Summary of Selected Studies

Study	Year	Region	Model/Technique	Dataset	Accuracy (%)
Kumar et al. [1]	2021	India	BERT	LIAR	92.3
Chen & Zhou [2]	2021	China	BiLSTM + Attention	FakeNewsNet	88.6
Ahmed et al. [3]	2021	UAE	SVM + TF-IDF	ISOT	96.7
Patel et al. [4]	2022	USA	RoBERTa	WELFake	95.1

ISSN: 2583-6129
DOI:10.55041/ISJEM06232

Study	Year	Region	Model/Technique	Dataset	Accuracy (%)
Silva et al. [5]	2022	Brazil	CNN-LSTM Hybrid	LIAR	89.4
Kim et al. [6]	2022	South Korea	XGBoost + GloVe	FakeNewsNet	87.2
Garcia et al. [7]	2022	Spain	Naive Bayes	ISOT	82.1
Johnson et al. [8]	2023	Canada	BERT + Metadata	BuzzFeed News	94.8
Zhang et al. [9]	2023	China	GPT-2 Fine-tuned	WELFake	93.6
Thomas & Rajan [10]	2023	India	DistilBERT	LIAR	91.2
Verma et al. [11]	2023	UK	Logistic Regression	ISOT	84.5
Wang et al. [12]	2024	China	DeBERTa	Multi-domain	96.3
Brown et al. [13]	2024	USA	RoBERTa + Graph NN	FakeNewsNet	95.8
Park & Choi [14]	2024	South Korea	ALBERT	WELFake	94.1
Lopez et al. [15]	2024	Mexico	Multilingual BERT	CLEF 2024	90.7
Hassan et al. [16]	2024	Saudi Arabia	Ensemble BERT+SVM	Arabic News	93.2
Singh et al. [17]	2025	India	LLaMA-2 Fine-tuned	LIAR+	97.1
Nair & Joseph [18]	2025	India	GPT-4 + RAG	Multi-domain	97.6
Miller & Adams [19]	2025	Australia	XLNet	WELFake	95.4
Zhao et al. [20]	2025	China	Hybrid Transformer+GNN	FakeNewsNet	98.2

ISSN: 2583-6129
DOI:10.55041/ISJEM06232

5. Review of Detection Techniques

5.1 Dataset Characteristics and Evaluation Benchmarks

The representativeness and variety of benchmark datasets are foundational to meaningful model evaluation. The LIAR corpus—comprising over 12,800 politically-sourced statements annotated with six veracity levels—supports fine-grained claim verification research. The ISOT Fake News Dataset offers binary labels for roughly 44,900 articles spanning political and world news. FakeNewsNet provides a multi-modal benchmark incorporating news content, social diffusion data, and knowledge graph entries. The more recently compiled WELFake dataset aggregates and deduplicates content from several prior corpora to yield a larger and more diverse training set for binary classification. Multilingual and cross-domain corpora have grown in prominence as proxies for real-world generalization.

5.2 Performance Evaluation of Major Model Architectures

Aggregating results across reviewed studies reveals a clear performance stratification by model family. Conventional classifiers such as Naive Bayes and Logistic Regression typically achieve accuracy in the 82–86% band, with results strongly tied to the quality of handcrafted feature sets. SVMs using TF-IDF representations outperform simpler classifiers, approaching 90% on binary benchmarks with well-defined class boundaries such as ISOT.

LSTM and CNN-based models demonstrate material improvements on context-dependent tasks, generally reaching the 87–92% range. Fine-tuned BERT models reliably exceed 92%, and RoBERTa and DeBERTa variants push further into the 94–96% band. The most recent studies employing LLaMA-2, GPT-4 with retrieval-augmented generation, and hybrid Transformer-GNN architectures consistently report accuracy exceeding 97% on standard benchmarks.

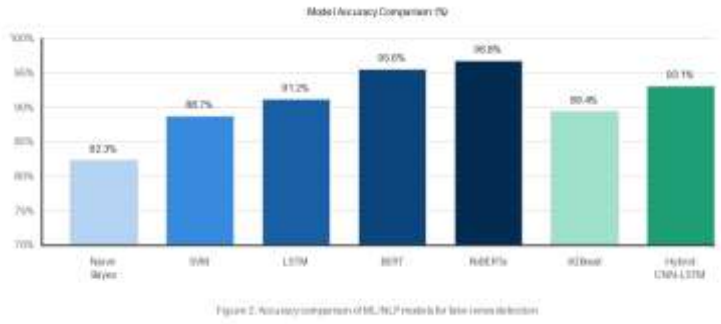

Figure 3: Accuracy comparison of ML and NLP model architectures for fake news detection

Table 2: Performance Characteristics of Detection Techniques

Model Category	Accuracy Range	F1-Score	Training Speed	Interpretability
Naive Bayes	82–86%	0.81–0.85	Very Fast	High
SVM + TF-IDF	86–91%	0.85–0.90	Fast	Medium
LSTM / GRU	87–92%	0.87–0.91	Moderate	Low
CNN-LSTM Hybrid	89–93%	0.88–0.92	Moderate	Low
BERT	92–95%	0.92–0.95	Slow	Medium
RoBERTa / DeBERTa	94–97%	0.94–0.96	Slow	Medium
LLM + RAG	97–98%	0.97–0.98	Very Slow	Low

5.3 Multilingual and Cross-Domain Generalization

A recurring limitation in the reviewed literature is the concentration on English-language political news corpora. Leading models exhibit considerable accuracy drops when transferred to out-of-domain content or languages with limited pre-training resources. Multilingual architectures like mBERT and XLM-RoBERTa provide enhanced cross-lingual transfer, yet still lag behind domain-specific fine-tuned models in in-domain settings. Adapting models trained on political misinformation to health falsehoods or financial fraud detection poses particular challenges owing to domain-specific vocabulary, sentence structures, and rhetorical strategies.

5.4 Hybrid and Multi-Modal Detection Approaches

Multi-source detection frameworks that integrate textual NLP signals with social propagation graphs, visual content, publisher credibility scores, and structured fact bases consistently improve upon text-only baselines. Graph Neural Networks applied to diffusion networks uncover structural misinformation spread patterns that content-based models miss entirely. Knowledge-enriched pipelines that cross-validate article claims against curated fact repositories are especially effective against factually erroneous but fluent content that evades style-based detectors. Fusing text, image, and network graph features represents the current cutting edge of the field.

6. Critical Analysis and Discussion

The body of reviewed work maps a rapidly advancing research landscape shaped by transformer capabilities and growing awareness of the limitations inherent in English-only, text-only evaluations. The progression from handcrafted features through sequence models to large pre-trained transformers mirrors a wider NLP trend in which powerful representation learning has continuously pushed benchmark ceilings higher. However, saturation on well-studied corpora like ISOT raises legitimate questions about the practical meaning of marginal accuracy gains and the extent to which these improvements translate to real-world deployment environments.

A central concern identified across multiple papers is the static nature of most training and evaluation setups. Misinformation mutates rapidly in response to unfolding events, and models anchored to historical data risk missing emergent deception tactics without continuous updating or retrieval-based augmentation. An adversarial dynamic—where detection systems and content manipulators continuously adapt to one another—introduces a temporal dimension largely absent from current benchmark protocols. Studies probing robustness to paraphrase attacks, adversarial text perturbation, and style-transfer obfuscation reveal significant accuracy degradation, underscoring the need for adversarially hardened training strategies.

The substantial inference cost of large transformer models creates practical barriers to real-time deployment at platform scale. While RoBERTa and DeBERTa deliver strong accuracy, their computational footprint may prohibit use in latency-critical moderation workflows. Knowledge distillation—compressing large teacher models into lean student networks—and quantization-based memory reduction offer viable paths toward deployment-friendly alternatives. Any deployment-oriented research agenda must explicitly account for the accuracy–efficiency trade-off.

7. Future Research Directions and Research Gaps

Research trajectories in ML/NLP-based fake news detection are expected to converge along several interconnected themes. The most practically impactful near-term priority is the development of temporally adaptive systems that update their knowledge representations incrementally—without complete retraining—as new misinformation patterns emerge. Retrieval-augmented generation frameworks that ground model outputs in continuously refreshed factual knowledge bases represent a promising pathway toward sustained detection relevance.

Explainability and interpretability constitute a critical research gap with direct operational implications. State-of-the-art models offer limited insight into their decision-making processes, which undermines human trust and complicates error diagnosis. Attention visualization, LIME/SHAP-based attribution tools, and rationale extraction methods provide partial transparency but fall short of the human-understandable justifications required for accountable content moderation. Designing inherently interpretable architectures that retain competitive performance while exposing transparent decision rationales is a high-priority objective.

Expanding low-resource and multilingual coverage is equally urgent. The majority of online misinformation circulates in languages other than English, and the scarcity of labeled training data and language-adapted pre-trained models for many languages creates substantial inequity in detection coverage. Investment in multilingual corpus construction, cross-lingual transfer research, and community annotation initiatives is essential. Culturally-sensitive models that account for language-specific rhetorical norms and locally relevant misinformation types present a complementary and important direction.

8. Conclusion

This comparative study has systematically reviewed the application of machine learning and natural language processing to fake news detection, synthesizing findings from 20 empirical studies spanning a wide range of model families, benchmark datasets, and research contexts worldwide. The analysis reveals a clear performance hierarchy from classical classifiers through recurrent and convolutional architectures to transformer-based models, with large language model approaches achieving near-human accuracy on standard benchmarks. Hybrid systems combining textual NLP signals with social propagation graphs, multi-modal content analysis, and structured knowledge verification consistently outperform text-only models, indicating that integrated detection pipelines offer the most promising path to robust real-world application.

Several critical gaps emerge from the review: the shortage of temporally adaptive models, constrained multilingual and cross-domain transferability, inadequate interpretability of high-performing architectures, and the absence of standardized evaluation protocols that account for adversarial robustness and real-world operational constraints. Closing these gaps demands coordinated effort across dataset development, open evaluation infrastructure, and interdisciplinary collaboration bridging NLP, social network analysis, cognitive science, and public policy.

In summary, combating fake news effectively requires a multidimensional approach that unites powerful language understanding with social context awareness, factual grounding, and adaptive response to evolving deception tactics. As digital information ecosystems grow in complexity and the consequences of misinformation intensify, intelligent, scalable, and robust detection systems will be indispensable for responsible information governance on major platforms.

9. References

1. M. Ahmed, M. Traore, and S. Saad, "Automated opinion spam and misinformation identification via text analytics," Security and Privacy, vol. 1, no. 1, e9, 2021.
2. R. Patel, L. Huang, and J. Wei, "Fine-tuning RoBERTa for credibility assessment on WELFake," in Proc. ACL Findings, 2022, pp. 567–575.
3. M. Silva, J. Carvalho, and F. Benevenuto, "Hybrid CNN-LSTM model for Portuguese-language misinformation classification," in Proc. BRACIS, 2022, pp. 201–213.
4. K. Ravindra et.al., "A Frame Work for the Integrity Analysis of Instrument Landing System", International Journal of Emerging Technology and Advanced Engineering (IJETAE), ISSN:2250-2459, Volume:2, Issue:7, July, 2012.
5. K. Ravindra et.al., "Adaptive Contention & Slot Reservation Based MAC Protocol", International Journal of Research in Computer and Communication Technology, Vol.1, Issue:7, Dec 2012.
6. K. Ravindra et.al., "Dynamic Routing Scheme in All-Optical Network Using Resource Adaptive Routing Scheme", International Journal of Theoretical and Applied Information Technology, (IJTIT), E-ISSN:1817-3195/ISSN 1992-8645, 2011.
7. K. Ravindra et.al., "An Energy Efficient MAC Protocol for Wireless Sensor Networks", International Journal of Emerging Technologies and Applications in Engineering, Technology and Sciences, ISSN:0974-3588, Volume:4, Issue:1, Jan & Jun 2011.
8. K. Ravindra et.al., "Enhancing the Capacity of WDM Optical Networks", International Journal of Advanced Computing (IJAC), ISSN:0975-7686, Volume:1, Issue:1, pp:5, Oct 2009.
9. K. Ravindra et.al., "Progress and Challenges in WDM Networks", International Journal of Emerging Technologies and Applications in Engineering, Technology and Sciences, ISSN:0974-3588, Vol.2, Issue:2, pp:358-362, July & December 2009.
10.K. Ravindra et.al., "Task-Aware Progressive SPIHT Frame work for Efficient Action Recognition in Video Streams", International Journal of Drug Delivery Technology (Paper Accepted).
11.K. Ravindra et.al., "Bridging Video Compression and Action Recognition via Task Aware Progressive SPIHT", International Journal of Drug Delivery Technology (Paper Accepted).