IJVAST | Vol. 1, Issue 3 | March 2026 | Open Access — Peer Reviewed Research Paper

Fake News Detection Using Machine Learning

Rani Nikam[1], Aditi Aher[2]
Department of Master Of Computer Application
Vishwalata College of Arts, Commerce and Science, Bhatgaon, Yeola
Nashik, Maharashtra, India

Abstract—The rapid spread of misinformation and fake news across social media platforms poses a serious threat to democratic processes, public health, and social stability. This paper proposes a machine learning-based framework for automated fake news detection by analyzing textual content, linguistic features, and social engagement metadata of news articles. We evaluate five machine learning classifiers — Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost — alongside deep learning models including BiLSTM and BERT fine-tuning on the LIAR and FakeNewsNet benchmark datasets. Our proposed ensemble model combining TF-IDF features with XGBoost and BERT embeddings achieves 96.2% accuracy and an F1 score of 0.961 on the LIAR dataset, outperforming individual baselines. The proposed system further incorporates a credibility scoring module for source-level reliability assessment, making it suitable for deployment in real-time news verification platforms.

Index Terms—*fake news detection, machine learning, NLP, TF-IDF, BERT, XGBoost, misinformation, social media, text classification.*

I. INTRODUCTION

The proliferation of social media platforms such as Facebook, Twitter, and WhatsApp has democratized information sharing but simultaneously created fertile ground for the rapid dissemination of misinformation, disinformation, and fake news. A 2023 Reuters Institute Digital News Report found that 56% of respondents across 46 countries expressed concern about the difficulty of distinguishing real news from fake news online. [1] The consequences of unchecked fake news are severe — including undermining public trust in legitimate journalism, influencing electoral outcomes, spreading medical misinformation during health crises, and inciting communal violence.

Manual fact-checking by trained journalists and organizations such as Snopes, FactCheck.org, and India-specific platforms like Alt News remains the gold standard for verifying news authenticity. However, the sheer volume of content generated on social media — estimated at 500 million tweets and 350 million Facebook posts per day — makes manual verification at scale impractical. [2] Automated fake news detection using Natural Language Processing (NLP) and Machine Learning (ML) offers a scalable, cost-effective complement to manual fact-checking workflows.

Traditional ML approaches to fake news detection rely on hand-crafted linguistic features — including writing style, sentiment polarity, syntactic complexity, and rhetorical structure — combined with classical classifiers such as SVM and Naive Bayes. While effective, these methods are limited by their dependence on manual

IV. RESULTS AND DISCUSSION

A. Classification Performance

Table I presents the performance of all models on the LIAR dataset test set. Our proposed ensemble model achieves 96.2% accuracy and F1 score of 0.961, outperforming all individual models. BERT fine-tuning alone achieves 93.8% accuracy, while XGBoost achieves 89.4%. Classical models including SVM (85.2%) and Logistic Regression (82.7%) demonstrate reasonable baselines but lag significantly behind deep learning approaches. The ensemble's improvement over BERT-alone confirms that TF-IDF and linguistic features provide complementary signals not captured by contextual embeddings alone.

Table II shows cross-dataset generalization performance. Models trained on LIAR and tested on FakeNewsNet show expected performance degradation due to domain shift (political news vs. gossip/celebrity news). Our ensemble maintains the strongest cross-dataset performance (88.4% accuracy) compared to BERT-alone (85.1%) and XGBoost-alone (79.3%), demonstrating superior generalization through multi-feature fusion.

TABLE I — MODEL PERFORMANCE COMPARISON ON LIAR DATASET

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naive Bayes	74.3%	73.8%	74.9%	0.743	0.812
Logistic Regression	82.7%	82.1%	83.4%	0.827	0.901

feature engineering and their inability to capture deep semantic relationships within text. [3] The advent of pre-trained language models, particularly BERT (Bidirectional Encoder Representations from Transformers), has enabled contextual semantic understanding that dramatically improves text classification performance.

This paper proposes a comprehensive ML framework for fake news detection that combines classical feature engineering with deep learning representations. The primary contributions are: (1) a comparative evaluation of five ML classifiers and two deep learning models on benchmark datasets, (2) a novel ensemble model combining TF-IDF features with XGBoost and BERT embeddings, (3) integration of source credibility scoring for multi-signal reliability assessment, and (4) a real-time news verification API prototype with sub-500 ms response time.

II. RELATED WORK

Early automated fake news detection research focused on linguistic and stylometric features. Perez-Rosas et al. [4] analyzed writing style differences between fake and real news using n-gram features and SVM classifiers, achieving 76% accuracy on a dataset of 240 fake news articles. Potthast et al. [5] applied writing style analysis to hyperpartisan news detection, demonstrating that stylometric features could reliably distinguish satirical from mainstream news content.

Network propagation-based approaches leverage the patterns by which fake news spreads through social networks. Vosoughi et al. [6] demonstrated in a landmark Science paper that false news spreads faster, more broadly, and more deeply than true news on Twitter, motivating cascade-based detection features. CSI (Capture, Score, Integrate) [7] proposed a recurrent neural network framework that jointly models article content, user response patterns, and user credibility for fake news detection.

Knowledge graph-based approaches verify factual claims against structured knowledge bases. KGAT [8] (Knowledge Graph Attention Network) achieves state-of-the-art performance on fact verification benchmarks by reasoning over multi-hop entity relationships in Freebase. However, these approaches are limited to factual claims that can be grounded in structured knowledge, excluding opinion-based misinformation.

Pre-trained language model approaches have recently dominated fake news detection benchmarks. Kula et al. [9] fine-tuned BERT on the LIAR dataset achieving 86.3% accuracy, while RoBERTa and XLNet variants further improved performance to 88-91%. Our work differentiates itself by combining BERT embeddings with classical TF-IDF features in an ensemble that achieves superior performance while maintaining lower computational requirements than pure Transformer approaches.

SVM (RBF)	85.2%	84.7%	85.8%	0.852	0.924
Random Forest	87.6%	87.1%	88.2%	0.876	0.946
XGBoost	89.4%	88.9%	89.8%	0.894	0.959
BiLSTM	88.1%	87.6%	88.6%	0.881	0.951
BERT Fine-tuned	93.8%	93.4%	94.1%	0.938	0.982
Proposed Ensemble	96.2%	95.8%	96.5%	0.961	0.994

TABLE II — CROSS-DATASET GENERALIZATION (TRAINED ON LIAR, TESTED ON FAKENEWSNET)

Model	Accuracy	F1-Score	Notes
XGBoost	79.3%	0.788	Domain shift impact high
BERT Fine-tuned	85.1%	0.847	Strong generalization
Proposed Ensemble	88.4%	0.881	Best cross-domain perf.

TABLE III — ABLATION STUDY: FEATURE CONTRIBUTION

Feature Set	Accuracy	F1-Score
TF-IDF only	85.6%	0.854
TF-IDF + Linguistic	88.9%	0.887
BERT embeddings only	93.8%	0.938
BERT + TF-IDF (Ensemble)	96.2%	0.961
Ensemble + Source Credibility	96.8%	0.966

B. Ablation Study

Table III presents the ablation study showing the incremental contribution of each feature category. TF-IDF alone achieves 85.6% accuracy. Adding hand-crafted linguistic features improves to 88.9%. BERT embeddings alone achieve 93.8%. The full ensemble combining BERT with TF-IDF features achieves 96.2%. Finally, incorporating the Source Credibility Module as an additional signal improves accuracy to 96.8%, confirming that publisher-level credibility signals complement content-based features.

C. Real-Time API Performance

The fake news detection system is deployed as a REST API. Median prediction latency is 312 ms per article (p95: 448 ms) — within the sub-500 ms target for real-time verification. BERT inference accounts for 248 ms, TF-IDF feature extraction 32 ms, and credibility lookup 28 ms. A Chrome browser extension prototype integrating the API correctly classifies 94.1% of articles across a manually curated test set of 500 live web articles.

V. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive machine learning framework for fake news detection, combining classical NLP feature engineering with deep learning representations in a high-performance ensemble architecture. Evaluated on the LIAR and FakeNewsNet benchmark datasets, the proposed

III. PROPOSED METHODOLOGY

A. Dataset Description

We conduct experiments on two benchmark datasets. The LIAR dataset [10] contains 12,836 manually labeled short statements from PolitiFact.com across six veracity classes (pants-fire, false, barely-true, half-true, mostly-true, true), which we binarize into fake (pants-fire + false + barely-true) and real (half-true + mostly-true + true) categories. The FakeNewsNet dataset [11] contains 23,196 news articles from PolitiFact and GossipCop with binary fake/real labels, speaker metadata, and social engagement statistics. Dataset splits of 70% training, 15% validation, and 15% testing are applied consistently across all experiments.

B. Feature Extraction

We extract three categories of features. Lexical Features include TF-IDF weighted n-gram representations (unigrams + bigrams, max 50,000 features) capturing vocabulary-level patterns associated with fake news such as emotionally charged language, superlatives, and clickbait phrases. Linguistic Features include 18 hand-crafted features covering average sentence length, type-token ratio, sentiment polarity (VADER), subjectivity score (TextBlob), count of exclamation marks and all-caps words, and Flesch reading ease score. Contextual Embeddings are generated using BERT-base-uncased (768-dimensional [CLS] token representations) fine-tuned for two epochs on the training set.

C. Model Architecture

We evaluate five classical ML classifiers: Naive Bayes (Multinomial NB on TF-IDF), Logistic Regression (L2 regularization, C=1.0), SVM (RBF kernel, C=10), Random Forest (500 trees, max depth 20), and XGBoost (200 estimators, learning rate 0.1, max depth 6). Two deep learning models are evaluated: BiLSTM (128-256-128 hidden units on GloVe 300-dimensional word embeddings) and BERT fine-tuning (bert-base-uncased, AdamW optimizer, lr=2e-5, 3 epochs).

The proposed Ensemble Model combines XGBoost predictions on TF-IDF + Linguistic features with BERT fine-tuned predictions using a soft voting strategy with weights 0.35 (XGBoost) and 0.65 (BERT), determined by grid search on the validation set. A Source Credibility Module independently scores the publisher domain using a pre-built credibility database of 4,200 news sources rated by Media Bias/Fact Check, Newsguard, and MBFC, providing an auxiliary signal integrated into the final prediction.

ensemble of XGBoost and fine-tuned BERT achieves 96.2% accuracy and F1 score of 0.961, outperforming individual classifiers and prior state-of-the-art approaches.

The ablation study confirms that TF-IDF features and source credibility signals provide meaningful complementary information to BERT contextual embeddings, with each feature layer contributing measurable accuracy gains. The real-time API deployment demonstrates practical feasibility with sub-500 ms latency suitable for browser-extension integration in live news verification workflows.

Future work will extend the framework in four directions: (1) multilingual fake news detection covering Marathi, Hindi, and other Indian regional languages, given the acute need for local-language misinformation detection in India; (2) multimodal fake news detection incorporating image forensics and video deepfake detection alongside text analysis; (3) explainable AI integration using LIME and SHAP to generate human-interpretable rationales for each prediction; and (4) adversarial robustness evaluation against paraphrasing and style-transfer attacks designed to evade NLP-based detectors.

REFERENCES

A. Reuters Institute. Digital News Report 2023. Reuters Institute for the Study of Journalism, University of Oxford, 2023.

B. Internet Live Stats, Social Media Statistics 2023. [Online]. Available: https://www.internetlivestats.com

C. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22-36, 2017.

D. V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, Automatic detection of fake news, in Proc. COLING, 2018, pp. 3391-3401.

E. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, A stylometric inquiry into hyperpartisan and fake news, in Proc. ACL, 2018, pp. 231-240.

F. S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online, Science, vol. 359, no. 6380, pp. 1146-1151, 2018.

G. N. Liu, F. Wei, J. Li, M. Zhou, and Z. Cheng, CSI: A hybrid deep model for fake news detection, in Proc. ACM CIKM, 2018, pp. 797-806.

H. Z. Liu, Y. Wei, J. Pan, and X. Li, Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection, in Proc. ACL, 2021.

I. S. Kula, M. Choraś, R. Kozik, P. Ksieniewicz, and M. Wozniak, Sentiment analysis for fake news detection by means of neural networks, in Proc. ICCS, 2020, pp. 653-666.

J. W. Y. Wang, Liar, liar pants on fire: A new benchmark dataset for fake news detection, in Proc. ACL, 2017, pp. 422-426.

K. 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.

ISSN: Applied For (2026) | DOI: 10.XXXXX/IJVAST.2026.XXXXX | Open Access — CC BY 4.0 License