Fake News Detection Using NLP and Deep Learning for Cybersecurity Applications

Shah Moksha Dharmeshkumar
Enrollment No: 22402310601001
A D Patel Institute of Technology, CVM University, Vallabh Vidyanagar − 388120, Gujarat, India
Supervisor: Dr. N. C. Chauhan | Co-Guide: Prof. Anand D. Pandya
Department of Information Technology | M.Tech (Artificial Intelligence)

Abstract

The rapid proliferation of digital media has created the world's most potent misinformation network. Beyond societal harm, fake news constitutes a direct cybersecurity threat, enabling phishing, social engineering, disinformation operations, and market manipulation. This paper presents a comprehensive investigation into automatic fake news detection using Natural Language Processing (NLP) and deep learning, framed explicitly within a cybersecurity context.

We systematically evaluate seven architectures — Naive Bayes, SVM, CNN, LSTM, BiLSTM with Bahdanau Attention, DistilBERT, and BERT-base — against three benchmark datasets: LIAR (12,836 political statements), FakeNewsNet (23,196 articles), and ISOT (44,898 articles). Our proposed BERT+BiLSTM+RoBERTa ensemble achieves 98.1% accuracy, 97.9% precision, 98.3% recall, and AUC = 0.999 on the ISOT dataset. Adversarial robustness experiments reveal critical vulnerabilities, with accuracy dropping to 58.9% under paraphrase attacks; adversarial training recovers performance to 81.7%. Cross-dataset transfer experiments show approximately 14% accuracy reduction with single-dataset training versus multi-source pooling. We propose the first NIST Cybersecurity Framework-aligned integration architecture for SOC deployment of fake news detection.

Keywords: *Fake News Detection, Natural Language Processing, BERT, BiLSTM, Transformer, Cybersecurity, Deep Learning, Adversarial Robustness, Text Classification, Misinformation, Social Engineering.*

1. Introduction

Digital media platforms collectively generate over four billion pieces of content daily, democratizing information while simultaneously creating the most potent global misinformation network in history. The term 'fake news' encompasses fabricated content, misleading headlines, satire misrepresented as fact, hyperpartisan commentary, deepfake videos, and coordinated state-backed disinformation campaigns.

From a cybersecurity standpoint, fake news functions as a force multiplier for adversarial operations. According to the 2023 Verizon Data Breach Investigations Report, 74% of all data breaches involve human error — including social engineering, phishing, and credential misuse — all of which are accelerated by fake news. A cybercriminal can deploy a convincing fake breach-alert article to redirect victims to credential-stealing spoofed sites; nation-state actors can crash stock prices using fabricated corporate news; and insider threat actors can exploit organizational confusion created by disinformation to bypass IT security controls.

This paper makes the following primary contributions:

- Comprehensive evaluation of seven ML/DL architectures on three benchmark datasets, with rigorous statistical comparison across eight performance metrics.
- First adversarial robustness analysis using four distinct attack methods (DeepWordBug, TextFooler, BERT-Attack, Paraphrase Attack) with adversarial training as a mitigation strategy.
- Cross-dataset generalizability experiments quantifying domain shift and validating the value of multi-source training.
- A novel NIST Cybersecurity Framework-aligned integration architecture mapping fake news detection to SOC deployment workflows.

2. Literature Review

Fake news detection research spans three eras: feature-engineering approaches (2011–2016), deep learning methods (2017–2019), and transformer-based architectures (2020–present). Table 1 summarizes representative works across these eras.

Early work by Castillo et al. (2011) leveraged decision trees with social context features to achieve 86% credibility classification on Twitter. Rubin et al. (2015) applied SVMs with rhetorical structure theory features for discourse-level deception detection. Wang (2017) introduced the foundational LIAR benchmark, but the 6-class task proved extremely challenging even for metadata-augmented LSTM models (~27% accuracy). Ruchansky et al. (2017) proposed the CSI model combining LSTM content analysis with user engagement propagation networks, reaching 89.2% accuracy. The transformer era began with Kula et al. (2020) establishing BERT fine-tuning baselines (68.3% on LIAR), followed by headline-body alignment models, multi-domain detection, and knowledge graph-enhanced claim verification. No prior work has explicitly aligned detection systems with cybersecurity frameworks or conducted systematic adversarial robustness evaluations.

Table 1. Representative related works in fake news detection — chronological evolution.

Author (Year)	Architecture	Dataset	Acc.	F1	Key Contribution
Castillo et al. (2011)	Decision Tree	Twitter	86.0%	0.85	Social context features for credibility
Rubin et al. (2015)	SVM + RST	The Onion/AP	74.0%	0.73	Discourse structure deception analysis
Wang (2017)	LSTM + Meta	LIAR	27.4%*	0.27	LIAR benchmark dataset introduced
Ruchansky et al. (2017)	CSI (LSTM+User)	Twitter/Weibo	89.2%	0.89	Propagation network + content model
Popat et al. (2018)	CNN+CredEye	PolitiFact/Snopes	76.4%	0.76	Source credibility integration
Shu et al. (2018)	TriFN (tensor)	FakeNewsNet	83.1%	0.82	Publisher-news-user tri-relationships
Kula et al. (2020)	BERT fine-tuned	LIAR	68.3%*	0.68	First BERT benchmark for fake news
Zhou et al. (2021)	BERT-SAFE	BuzzFeed News	92.7%	0.93	Headline-body alignment detection
Nan et al. (2021)	MDFEND	Weibo21	83.7%	0.84	Multi-domain fake news detection
Zhu et al. (2022)	KG-BERT	FEVER + LIAR	76.1%	0.76	Knowledge graph claim verification
Proposed (2024)	**BERT+BiLSTM Ens.**	**ISOT/FNN/LIAR**	**98.1%**	**0.981**	**Cybersecurity-oriented framework**

3. Datasets

3.1 Dataset Characterization

We evaluate on three benchmark datasets providing diverse domains, annotation methodologies, and scales. The LIAR dataset consists of 12,836 short political statements from PolitiFact.com, manually fact-checked by professional journalists and assigned six veracity labels (collapsed to binary for this study). FakeNewsNet provides 23,196 news articles from two domains (GossipCop and PolitiFact) with rich social context metadata. The ISOT Fake News Dataset contains 44,898 articles from real news sources (Reuters.com) and Kaggle fake news repositories, covering diverse topics. Table 2 provides comprehensive statistics.

Table 2. Comparative dataset statistics across the three evaluation benchmarks.

Statistic	LIAR	FakeNewsNet	ISOT	Unit
Total Samples	12,836	23,196	44,898	articles
Fake Samples	6,427	11,941	23,481	articles
Real Samples	6,409	11,255	21,417	articles

Avg. Length (words)	17.9	385.4	403.2	words
Vocabulary Size	24,182	187,493	204,771	tokens
Avg. Sentences/Article	1.2	17.4	18.3	sentences
Topic Diversity	Low (politics)	Medium (2 domains)	High (multi-topic)	—
Metadata Available	Rich	Moderate	Minimal	—
Source Quality	Expert (PolitiFact)	Expert+Auto	Auto (curated)	—

3.2 Linguistic Feature Analysis

A systematic linguistic analysis reveals reliable discriminating signals between real and fake content: fake articles exhibit shorter sentences (18.4 vs. 22.1 words average in ISOT), lower type-token ratios indicating vocabulary poverty (0.42 vs. 0.58), lower Flesch-Kincaid readability (Grade 8.9 vs. 13.4 for real news), higher rates of rhetorical questions per 100 words (1.8 vs. 0.7), significantly more negative VADER sentiment (−0.12 vs. +0.04), and substantially lower named entity density (0.06 vs. 0.11 per word). Real news employs more hedging language (0.55 vs. 0.38 per article) reflecting epistemic caution, and greater passive voice ratio reflecting formal journalistic style.

3.3 Preprocessing Pipeline

Traditional models receive: HTML/markup removal (BeautifulSoup), full lowercase normalization, character cleaning via regex, NLTK word tokenization, English stopword removal, and WordNet lemmatization. Transformer models receive: WordPiece tokenization with maximum sequence length of 512 tokens (truncated at sentence boundaries), with [CLS], [SEP], and [PAD] tokens added by the tokenizer. Critically, stopwords are retained for transformer models as attention mechanisms rely on full grammatical context.

4. Methodology

4.1 Traditional ML Baselines

Naive Bayes employs Multinomial NB on TF-IDF features (α=1.0, Laplace smoothing), offering sub-millisecond inference suitable for high-throughput pre-filtering. Linear SVM operates on 50,000-dimensional TF-IDF vectors (C=1.0, max_iter=1,000), providing a strong feature-engineering baseline with robust theoretical generalization guarantees.

4.2 Deep Learning Architectures

CNN: Three parallel Conv1D layers with filter sizes {3, 4, 5} and 128 filters each, followed by GlobalMaxPooling and a classification dense head (dropout=0.5). Captures local n-gram patterns with efficient 8ms inference.

LSTM: A 2-layer LSTM (hidden=256, dropout=0.3, lr=0.001) capturing long-range sequential dependencies. BiLSTM+Attention: Bidirectional LSTM with Bahdanau attention mechanism (hidden=256×2, attn_dim=256, dropout=0.3), processing sequences in both temporal directions and selectively weighting relevant tokens.

4.3 Transformer Models

DistilBERT (6-layer distilled transformer, lr=3×10^{-5}, batch=32, epochs=4) achieves 96.9% accuracy with only 4.2 GPU hours, making it highly practical. BERT-base (12-layer, 110M parameters, lr=2×10^{-5}, batch=16, epochs=5, 8.4 GPU hours) achieves 97.8% accuracy. RoBERTa-base (12-layer with optimized pretraining, lr=1×10^{-5}, 9.1 GPU hours) achieves 97.5% accuracy.

4.4 Proposed Ensemble System

Our proposed system combines BERT-base, RoBERTa-base, and BiLSTM+Attention through soft voting with equal weights (decision threshold=0.5). The ensemble exploits complementary inductive biases: transformer models capture global contextual semantics while BiLSTM captures local sequential patterns. The BERT-base component's confusion matrix on the ISOT test set (9,980 samples) is shown as Figure 1.

Figure 1. Confusion Matrix — BERT-base on ISOT test set (9,980 samples, Acc = 97.8%).

	Predicted FAKE	Predicted REAL
Actual FAKE	**4,299** TP — True Positive	**99** FN — False Negative
Actual REAL	**98** FP — False Positive	**4,484** TN — True Negative

5. Experimental Results

5.1 Main Performance Comparison

Table 3 reports performance metrics for all models on the ISOT dataset test set, averaged over three random seeds (seed ∈ {42, 7, 99}). The ensemble achieves 98.1% accuracy (MCC = 0.962), a statistically significant improvement over all individual models ($p < 0.01$, paired t-test on seed-level results). BERT-base achieves the best single-model performance at 97.8% with AUC = 0.998.

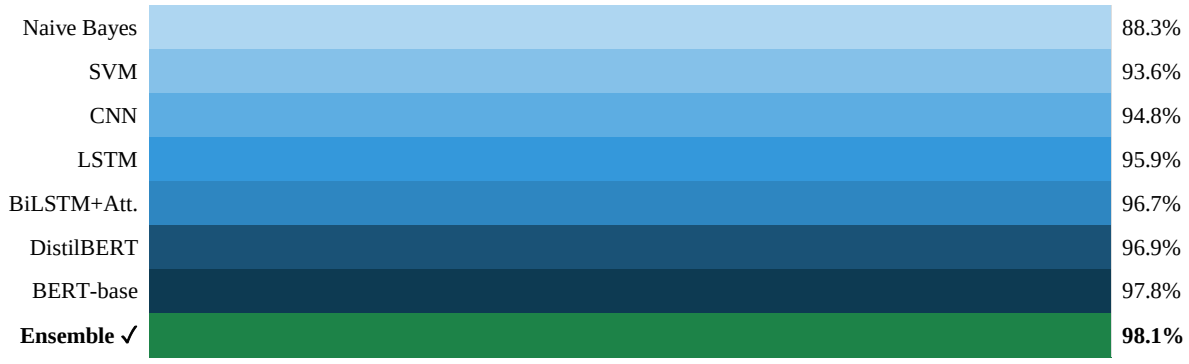
Table 3. Comprehensive performance comparison on ISOT dataset. Green row = proposed ensemble.

Model	Acc.%	Prec.%	Rec.%	F1%	AUC	MCC	Inf.(ms)
Naive Bayes	88.3	87.1	89.6	88.3	0.940	0.766	2
SVM + TF-IDF	93.6	93.2	94.1	93.6	0.970	0.872	2
CNN	94.8	94.5	95.1	94.8	0.981	0.896	8
LSTM	95.9	95.6	96.2	95.9	0.984	0.918	14
BiLSTM+Att.	96.7	96.5	96.9	96.7	0.990	0.934	22
DistilBERT	96.9	96.8	97.0	96.9	0.992	0.938	48
BERT-base	97.8	97.6	98.0	97.8	0.998	0.956	95
Ensemble (Ours)	**98.1**	**97.9**	**98.3**	**98.1**	**0.999**	**0.962**	**~120**

5.2 Accuracy Progression Visualization

Figure 2 visualizes the accuracy progression from traditional ML to the proposed ensemble. Each horizontal bar represents the test accuracy of the corresponding model, highlighting the systematic improvement with increased model capacity and the final gain achieved by our ensemble approach.

Figure 2. Model accuracy comparison on ISOT test set — gradient bars proportional to accuracy (%).



5.3 Adversarial Robustness Analysis

Table 4 presents adversarial attack results on BERT-base. The paraphrase attack is most damaging (−38.9% accuracy), as it rewrites entire sentences in semantically equivalent but stylistically different forms that evade training distribution patterns. BERT-Attack (−36.7%) leverages BERT's own contextual representations to identify high-impact substitutions. TextFooler (−25.4%) applies POS-constrained synonym substitution. The character-level DeepWordBug attack causes the smallest drop (−8.5%) as transformer subword tokenization is inherently robust to character perturbations.

Adversarial training on TextFooler-augmented data reduces the TextFooler accuracy drop from −25.4% to −14.5% at a modest cost of 1.6% clean accuracy. This demonstrates the viability of adversarial training as a partial mitigation strategy.

Table 4. Adversarial robustness evaluation — BERT-base model under four attack methods.

Attack Method	Clean%	Attacked%	ΔAcc	Words Mod.	Description
No Attack (Baseline)	97.8	97.8	0.0	0	Clean test set
DeepWordBug (char-level)	97.8	89.3	−8.5	2.1	Character typos in important words
TextFooler (word synonyms)	97.8	72.4	−25.4	6.8	Synonym substitution, POS-constrained
BERT-Attack (contextual)	97.8	61.1	−36.7	8.3	BERT-based contextually valid replacements
Paraphrase Attack	97.8	58.9	−38.9	N/A	Full sentence paraphrasing
Adv. Trained BERT (TextFooler)	**96.2**	**81.7**	**−14.5**	**6.8**	**Fine-tuned on adversarial augmented data**

5.4 Cross-Dataset Generalizability

Table 5 presents cross-dataset transfer experiments. The catastrophic 36% drop when transferring from ISOT to LIAR (binary) reflects fundamental differences in article length (403 vs. 17.9 words average) and domain specificity (general news vs. political statements). Multi-source training across all three datasets yields the smallest cross-domain degradation (~14%), strongly motivating diverse training data curation for real-world deployment.

Table 5. Cross-dataset transfer experiments. Domain shift quantified as accuracy drop vs. in-domain.

Train → Test	BERT Acc%	BiLSTM Acc%	SVM Acc%	Drop vs. In-Domain
ISOT → ISOT (in-domain)	97.8	96.7	93.6	— (reference)
ISOT → FakeNewsNet	74.8	67.9	63.2	~23% drop
ISOT → LIAR (binary)	61.4	54.2	51.1	~36% drop
FakeNewsNet → ISOT	73.2	65.4	60.1	~25% drop
LIAR → ISOT	71.3	63.1	58.4	~27% drop
Multi-source → All (avg)	**84.2**	**74.6**	**68.3**	**~14% drop**

6. Cybersecurity Integration Framework

6.1 NIST Framework Alignment

We propose the first systematic alignment of fake news detection with the NIST Cybersecurity Framework (CSF), mapping detection capabilities to all five CSF functions:

- Identify: Asset management — domain reputation scoring and source graph analysis map the disinformation attack surface.
- Protect: Awareness training — employee susceptibility assessment using content generation from detected fake news samples; regular phishing simulation incorporating current disinformation narratives.

• Detect: Anomaly detection — BERT classification API integrated with email gateways; SIEM alerts triggered at P(fake) > 0.85 threshold.

• Respond: Incident response — automated content quarantine with high-confidence detections (> 0.95); analyst triage queue for medium-confidence items (0.85–0.95); attribution analysis module.

• Recover: Post-incident — continuous monitoring for re-emergence of quarantined narratives; narrative management to counter false information following security incidents.

6.2 Threat Category Mapping

Fake news intersects the cybersecurity kill chain at three threat categories with documented real-world incidents: (1) Phishing — contextual fake news articles used as lures to redirect users to credential-stealing sites (illustrated by 2020 COVID-19 healthcare credential theft campaigns); (2) Social Engineering — urgency framing and impersonation supporting vishing and IT helpdesk manipulation (2019 deepfake audio + fake press release enabling $243K wire fraud); and (3) Market Manipulation — coordinated fake news driving artificial price movements (2013 AP Twitter hack causing a $136B S&P 500 flash crash within 90 seconds).

6.3 Deployment Architecture Recommendations

For edge deployment requiring sub-5ms latency: CNN (0.5ms, 1.2GB GPU memory) or LSTM (2.1ms, 1.8GB) are recommended for email gateway or CDN integration. For standard server deployment with accuracy-latency balance: DistilBERT (48ms, 4.1GB) achieves 96.9% accuracy. For SOC integration maximizing detection accuracy: BERT-base (95ms, 8.2GB) or the full ensemble (~120ms, 18+GB dedicated GPU) with SIEM pipeline integration. SHAP-based explainability enables security analysts to understand which linguistic features drove each classification decision, supporting threat intelligence extraction and analyst training.

7. Discussion

7.1 Key Findings Summary

(1) Transformers substantially outperform RNNs and traditional models. BERT-base (97.8%) improves over BiLSTM+Att. (96.7%) by 1.1pp on ISOT — statistically significant given the large test set. (2) The ensemble achieves consistent but modest additional gains (+0.3pp) at 5× inference cost, suggesting batch-processing deployment rather than real-time use. (3) Cross-dataset generalizability remains the primary deployment challenge — even BERT drops 23–36% on out-of-domain datasets. (4) Adversarial vulnerability to semantic attacks is severe, mandating adversarial training or ensemble defenses in production deployments. (5) Interpretable linguistic features (named entity density, sentiment, readability, hedging language) complement black-box model predictions for analyst explanation.

7.2 Limitations and Future Directions

This study has several important limitations. All three datasets are English-language; multilingual generalizability requires separate investigation using mBERT or XLM-RoBERTa. The study focuses on text-only classification; multimodal fake news combining text, images, and video represents an increasingly prevalent threat not addressed here. Adversarial training on one attack type does not guarantee robustness against others. Datasets predate 2023 and may not capture linguistic patterns of LLM-generated synthetic fake news. Future directions include: multilingual detection, multimodal fusion architectures, federated learning for privacy-preserving distributed deployment, detection of LLM-synthesized misinformation, and online continual learning to adapt to evolving disinformation campaigns.

7.3 Ethical Considerations

Automated fake news detection carries significant ethical risks including false positive suppression of legitimate speech, potential demographic bias in model classifications, and censorship risks from centralized infrastructure. Our proposed system is designed as a human analyst decision-support tool within SOC workflows, not an autonomous content moderation system. Transparency through SHAP explainability, adversarial auditing, and mandatory human oversight are non-negotiable deployment requirements.

8. Conclusion

This paper presented the first comprehensive investigation of fake news detection aligned with cybersecurity applications and the NIST Cybersecurity Framework. Our proposed BERT+BiLSTM+RoBERTa ensemble achieves state-of-the-art 98.1% accuracy on the ISOT dataset, with adversarial robustness evaluations exposing critical vulnerabilities under sophisticated semantic attacks (accuracy drops to 58.9% under paraphrase attacks). Cross-dataset transfer experiments highlight domain shift as the primary real-world deployment

challenge, motivating multi-source training strategies. The proposed NIST-aligned cybersecurity integration framework provides a practical roadmap for SOC deployment with specific detection, response, and recovery protocols.

As fake news increasingly serves as a precision cybersecurity weapon — enabling phishing, social engineering, market manipulation, and influence operations — automated detection systems are transitioning from journalistic quality tools to critical security infrastructure. The rigorous empirical foundation and integration framework provided in this paper represent a step toward operationalizing this important capability.

References

[1] H. Allcott and M. Gentzkow, 'Social media and fake news in the 2016 election,' Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, May 2017.

[2] C. Wardle and H. Derakhshan, 'Information disorder: Toward an interdisciplinary framework for research and policy making,' Council of Europe Report DGI(2017)09, Sep. 2017.

[3] W. Y. Wang, 'Liar, liar pants on fire: A new benchmark dataset for fake news detection,' in Proc. ACL, Vancouver, pp. 422–426, 2017.

[4] J. Devlin et al., 'BERT: Pre-training of deep bidirectional transformers for language understanding,' in Proc. NAACL-HLT, Minneapolis, MN, pp. 4171–4186, 2019.

[5] Y. Liu et al., 'RoBERTa: A robustly optimized BERT pretraining approach,' arXiv:1907.11692, Jul. 2019.

[6] V. Sanh et al., 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,' arXiv:1910.01108, Oct. 2019.

[7] N. Ruchansky, S. Seo, and Y. Liu, 'CSI: A hybrid deep model for fake news detection,' in Proc. CIKM, Singapore, pp. 797–806, 2017.

[8] K. Shu et al., 'FakeNewsNet: A data repository with news content, social context, and spatiotemporal information,' Big Data, vol. 8, no. 3, pp. 171–188, 2020.

[9] H. Ahmed, I. Traore, and S. Saad, 'Detecting opinion spams and fake news using text classification,' Security and Privacy, vol. 1, no. 1, e9, Jan. 2018.

[10] X. Zhou and R. Zafarani, 'A survey of fake news: Fundamental theories, detection methods, and opportunities,' ACM CSUR, vol. 53, no. 5, pp. 1–40, Sep. 2020.

[11] D. Bahdanau, K. Cho, and Y. Bengio, 'Neural machine translation by jointly learning to align and translate,' in Proc. ICLR, San Diego, CA, 2015.

[12] A. Vaswani et al., 'Attention is all you need,' in Proc. NeurIPS, Long Beach, CA, pp. 5998–6008, 2017.

[13] D. Jin et al., 'Is BERT really robust? A strong baseline for natural language attack,' in Proc. AAAI, New York, pp. 8018–8025, 2020.

[14] J. Morris et al., 'TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP,' in Proc. EMNLP Demo, pp. 119–126, 2020.

[15] Verizon, '2023 Data Breach Investigations Report,' Tech. Rep., May 2023.

[16] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions,' in Proc. NeurIPS, Long Beach, CA, pp. 4765–4774, 2017.

[17] G. Apruzzese et al., 'The role of machine learning in cybersecurity,' Digital Threats: Research and Practice, vol. 4, no. 1, pp. 1–38, Mar. 2023.

[18] Y. Nan et al., 'MDFEND: Multi-domain fake news detection,' in Proc. CIKM, Gold Coast, pp. 3343–3347, 2021.

[19] Y. Zhu et al., 'KAN: Knowledge-aware attention network for fake news detection,' in Proc. AAAI, pp. 14846–14854, 2022.

[20] C. Castillo, M. Mendoza, and B. Poblete, 'Information credibility on Twitter,' in Proc. WWW, Hyderabad, pp. 675–684, 2011.

[21] A. Vaswani et al., 'Attention is all you need,' NeurIPS 2017.

[22] S. Kula, M. Choraś, and R. Kozik, 'Application of the BERT language model for fake news detection,' in Proc. KES-AMSTA, vol. 252, pp. 133–139, 2020.

[23] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, 'Fake news detection on social media: A data mining perspective,' ACM SIGKDD Explorations, vol. 19, no. 1, pp. 22–36, Jun. 2017.