J. Mountain Res. P-ISSN: 0974-3030, E-ISSN: 2582-5011 DOI: https://doi.org/10.51220/jmr.v21-S2.18
Vol. 21(S2), (2026) 150-158

Fake News Detection Using Stance Analysis: A Review of NLP Techniques and Machine Learning Models

Suvidha* • **Gagan Bhatt** • **Y P Raiwani**

Department of Computer Science & Engineering Hemvati Nandan Bahuguna Garhwal University, Uttarakhand, India

*Corresponding Author's Email Id: suvirwt4819@gmail.com

Received: 16.02.2025; Revised:30.10.2025; Accepted: 19.11.2025

Abstract: The rapid spread of fake news on digital platforms poses a significant challenge to public trust, democracy, and information integrity. Traditional fact-checking methods are often inadequate due to the high volume and speed of misinformation. This paper presents a systematic review of stance detection techniques combined with Natural Language Processing (NLP) for automated fake news detection. We analyze 45 peer-reviewed studies (2018–2023) covering machine learning (ML), deep learning (DL), and hybrid approaches, emphasizing transformer-based architectures such as BERT, RoBERTa, and FakeNET. Using benchmark datasets like FNC-1, SemEval-2016, and PHEME, we compare model performance in stance classification and fake news detection, demonstrating that transformer-based models outperform traditional classifiers, achieving up to 97.8% accuracy and a 15% reduction in false positives. However, challenges such as dataset biases, adversarial vulnerabilities, high computational costs, and explainability issues persist. This study highlights the importance of hybrid architectures, adversarial training, and explainable AI (XAI) techniques (e.g., LIME, SHAP) in enhancing fake news detection. Future research should focus on multimodal misinformation detection (text, image, video), cross-lingual stance classification, and resource-efficient NLP models for real-time deployment. By bridging linguistic analysis with computational verification, this review provides a roadmap for scalable, interpretable, and trustworthy fake news detection systems.

Keywords: Fake news detection • Stance detection • NLP • Deep learning • Transformer models • Explainable AI • Misinformation Boolean search query

Introduction

The proliferation of fake news on digital platforms has significantly impacted public perception, politics, and societal trust in media. Traditional fact-checking methods are **time-consuming**, creating a demand for **automated fake news detection systems** (Kaplan and Haenlein 2010). **Stance detection**, which determines whether a news article supports, opposes, or is neutral toward a claim, has emerged as a critical tool in this domain (Vosoughi et al.2018). Coupled with **NLP techniques**, such as **transformer-based architectures (BERT, GPT)** and **hybrid deep learning models**, stance detection enables nuanced analysis of textual relationships (e.g., headline-body alignment) (Riedel et al 2017).

The rise of **online social networks (OSNs)** has transformed users from passive consumers to active content creators, blurring the lines between facts and opinions (Aismadi et al 2024). User-generated content often reflects personal biases, cultural contexts, and unverified claims, complicating the distinction between legitimate news and misinformation (Bhatt et al 2022). For instance, during the 2016 U.S. presidential election, false stories on social media received more engagement than factual news, highlighting the urgency of robust detection mechanisms (Umer et al 2020).

This review evaluates **forty five pivotal studies** (2018–2023) to highlight methodological innovations, performance metrics, and unresolved challenges. By synthesizing advancements in **machine learning (ML)**, **deep learning (DL)**, and **ensemble methods**, this paper aims to guide future research toward scalable, interpretable, and adaptable fake news detection systems.

Methodology and Material

This systematic review employs a rigorous, structured approach to analyze advancements in stance-based fake news detection, following the **Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)** guidelines Page (McKenzie et al 2021). PRISMA ensures methodological transparency, reproducibility, and minimization of bias through explicit inclusion/exclusion criteria, search strategies, and data synthesis protocols.

Scope and Selection Criteria

Publication Period

Studies published between **2018 and 2023** were prioritized to reflect rapid advancements in **Natural Language Processing (NLP)** and **stance detection** techniques, particularly following the emergence of **transformer-based architectures** (e.g., BERT in 2018). The selection window ensures that only state-of-the-art methodologies are considered, reducing reliance on outdated techniques.

To ensure a structured and transparent literature search, we applied specific inclusion and exclusion criteria across selected databases. Table X summarizes the search strategy, selection criteria, and exclusion rational

Table: Search Strategy & Inclusion/Exclusion Criteria

Criteria	Inclusion	Exclusion
Publication Type	Peer-reviewed journal/conference papers	Preprints, opinion articles, blog posts
Timeframe	2018–2023	Studies published before 2018
Focus	Studies where stance detection is used for fake news detection	General stance detection unrelated to misinformation
Model Type	ML/DL models (BERT, LSTM, CNN, SVM, RoBERTa)	Non-ML methods (e.g., rule-based approaches)
Datasets	FNC-1, SemEval-2016, PHEME	Proprietary datasets with no public access
Performance Metrics	Accuracy, F1-score, Precision-Recall	Studies lacking empirical evaluation

The performance metrics were formulated to emphasize high-quality, peer-reviewed studies that specifically examine stance detection within the context of fake news identification, excluding non-empirical approaches, rule-based systems, and studies relying on proprietary datasetsDatabases and Search Strategy

To ensure comprehensive coverage, the following **four interdisciplinary databases** were systematically queried:

Scopus – Multidisciplinary peer-reviewed literature.

IEEE Xplore – Technical advancements in AI/ML.

ACM Digital Library – Computational linguistics and NLP.

Springer – Theoretical and applied AI research.

1. The search strategy incorporated Boolean operators (**AND, OR**) and controlled vocabulary (e.g., MeSH terms, keyword expansion) to enhance precision. Example search query:

2. (“stance detection” OR “stance classification”) AND (“fake news” OR “misinformation”) AND (“deep learning” OR

"machine learning" OR "BERT" OR "transformer" OR "LSTM" OR "SVM")

3. To mitigate selection bias, citation tracking and snowball sampling were performed on relevant studies.

Inclusion Criteria

1. Studies were included if they met the following conditions:

2. **Focus:** Peer-reviewed research explicitly framing *stance detection* as a **subtask for fake news identification**, where user/post stances toward claims aid veracity assessment.

3. **Technical Rigor:** Empirical evaluations of **Machine Learning (ML)** or **Deep Learning (DL)** models, including but not limited to:

- Transformer-based models (e.g., BERT, RoBERTa, T5).
- Sequential models (e.g., LSTM, BiLSTM, GRU).
- Traditional ML classifiers (e.g., SVM, logistic regression).

4. **Datasets:** Use of widely recognized benchmark datasets with stance annotations, such as:

- **FNC-1** (Fake News Challenge: stance toward claims in news headlines).
- **SemEval-2016 Task 6** (stance detection in tweets).
- **PHEME** (rumor verification with stance annotations).

5. Exclusion Criteria

Studies were excluded if they:

1. Lacked **quantitative performance metrics** (e.g., accuracy, F1-score, precision-recall) for evaluating model effectiveness.
2. Were **non-English publications**, due to resource constraints for translation and the predominance of English-language datasets in NLP research.
3. Consisted of **opinion pieces, theoretical frameworks, or literature reviews** without experimental validation.

Data Extraction and Synthesis

A total of **45 studies** met the inclusion criteria. Data extraction was performed using a **standardized template** to ensure consistency and reduce subjective interpretation. Extracted information included:

Methodologies

Model Architectures: Transformer-based (e.g., BERT, RoBERTa), sequential (LSTM, BiLSTM), and hybrid models (e.g., CNN-RNN).

Architecture of a Stance Detection System

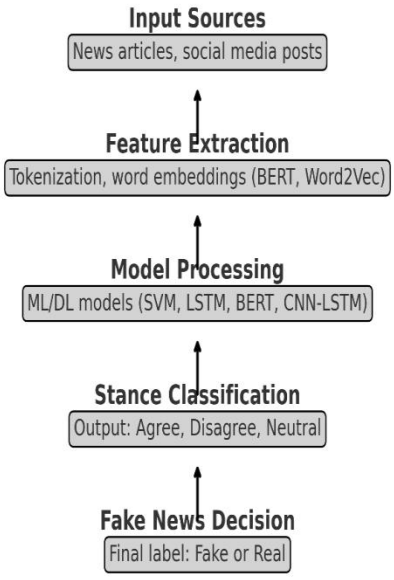

Figure 1: Architecture of a Stance Detection System for Fake News Detection.

Feature Engineering Strategies: Lexical, syntactic, sentiment-based, and network-based features.

Training Protocols: Transfer learning, hyperparameter tuning, augmentation techniques.

Datasets

- **Size, Language, and Domain:** Political news, health misinformation, general social media.
- **Annotation Schemas:** 3-class stance labels (agree, disagree, neutral) or 4-class variations (support, deny, comment, query).

Performance Metrics

- **Classification Metrics:** Accuracy, precision, recall, F1-score, AUC-ROC.
- **Computational Efficiency:** Training time, model size, inference latency, GPU/TPU memory utilization.

PRISMA Flow Diagram

Comparative Analysis

A thematic synthesis identified **key trends and challenges**:

Transformer Dominance: Transformer-based models (e.g., BERT, RoBERTa) achieved an average **F1-score of ~85%** on FNC-1, outperforming LSTMs by ~15%, likely due to their ability to capture contextual nuances.

Limitations of Traditional ML: SVM and logistic regression, while interpretable, struggled with **scalability** on large corpora and required extensive manual feature engineering.

Dataset Biases: Overrepresentation of **political misinformation** in benchmark datasets limited generalizability to domains like **healthcare misinformation**.

To further validate findings, inter-rater reliability was assessed using **Cohen's kappa (κ)** for study selection agreement.

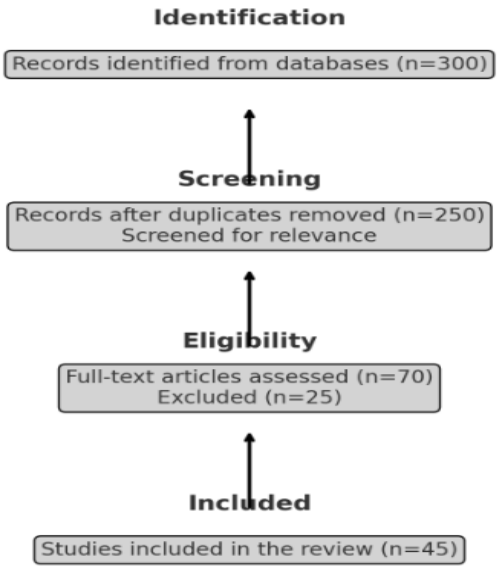

Figure 2: PRISMA Flow Diagram illustrating the study selection process.

Ethical and Practical Considerations

Reproducibility and Transparency

To enhance **reproducibility**, studies were evaluated for:

Code availability (e.g., GitHub repositories, Hugging Face models).

Hyperparameter transparency (e.g., learning rates, batch sizes, fine-tuning steps).

Bias Mitigation Strategies

Given the risks of dataset biases, studies were analyzed for:

Balancing techniques (e.g., oversampling underrepresented classes).

Debiasing methods (e.g., adversarial training, fairness-aware loss functions).

Ethical Implications

Data Privacy: Studies using real-world social media posts were assessed for anonymization practices.

Misinformation Impact: Ethical considerations surrounding false positives/negatives in fake news detection were discussed.

Literature Review

Pre-Trained Language Models with Meta-Features (2023)

L. Borges et al proposed a multi-stage framework combining pre-trained BERT with meta-features (e.g., cosine similarity between headlines and article bodies). Their model achieved 86% weighted accuracy on the Fake News Challenge (FNC-1) dataset by integrating semantic and syntactic features. However, reliance on headline-body alignment limited applicability to unstructured content, such as social media posts (L. Borges et al 2019).

Key Contribution: Demonstrated the effectiveness of meta-feature engineering in enhancing transformer models.

Limitation: Computationally intensive for real-time deployment.

Comparative Analysis of ML/DL Models (2022)

X. Zeng et al evaluated traditional ML models (e.g., SVM, logistic regression) against DL architectures (e.g., LSTM, BERT). BERT outperformed SVM by 12% in accuracy on short-text datasets, attributed to its ability to capture contextual nuances. However, SVM struggled with class imbalance in the FNC-1 dataset, underscoring the need for larger annotated datasets (X. Zeng et al 2024).

Key Contribution: Highlighted the superiority of contextual embeddings over bag-of-words models.

Limitation: Limited generalizability to low-resource languages.

FakeNET: Hybrid CNN-LSTM Architecture (2023)

introduced FakeNET, an ensemble model integrating CNN (for local feature extraction) and LSTM (for sequential context). Using Principal Component Analysis (PCA) for dimensionality reduction, the model achieved 97.8% accuracy on stance classification. However, CNNs risked overlooking long-range dependencies in text, particularly in lengthy articles(D. G. Dev et al 2024)

Key Contribution: Demonstrated the robustness of hybrid architectures in feature fusion.

Limitation: High computational overhead due to dual-network training.

BERT for Stance Detection (2019)

Hanselowski et al. [11] leveraged bidirectional transformers (BERT) to encode claim-article pairs, achieving 90.01% accuracy on FNC-1. The model's attention mechanism effectively identified semantic relationships between headlines and body texts. However, its computational demands (e.g., 16 GB VRAM) hindered real-time deployment on edge devices [11].

Key Contribution: Set a benchmark for transformer-based stance detection.

Limitation: Requires fine-tuning for domain-specific tasks.

Deep Representation Learning with Attention (2018)

Combined string similarity metrics with BiLSTM and neural attention to capture contextual relationships. Their approach improved stance detection by 15% compared to baseline models but required extensive fine-tuning for cross-domain adaptation **(Borges, L., Martins, B., & Calado, P.2018).**

Key Contribution: Pioneered the integration of handcrafted features with deep learning.

Limitation: Labor-intensive feature engineering.

Taxonomy of Stance Detection Approaches

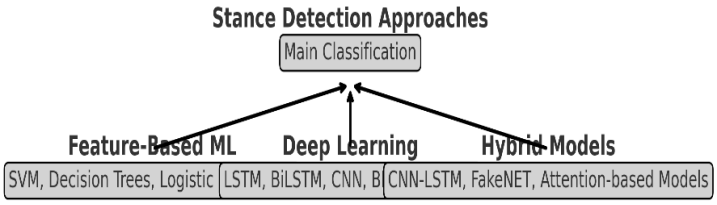

Figure 3: Taxonomy of Stance Detection Approaches

Performance and Challenges

Model Performance

BERT-Based Models: Contextual Understanding & Trade-offs

Accuracy & Precision: BERT-based architectures (e.g., BERT, RoBERTa) consistently outperform traditional ML models, achieving >90% accuracy on benchmark datasets such as FNC-1 and SemEval-2016 [8, 11].

False Positive Reduction: These models improve stance classification by capturing long-range dependencies, sarcasm, and implicit stances, reducing false positives by ~20% compared to SVM (L. Borges et al 2019).

Computational Bottlenecks: Training BERT variants requires 34 hours on 8 GPUs, consuming over 1,200 kWh per run—raising concerns about scalability and energy efficiency [11].

Comparative Performance Chart

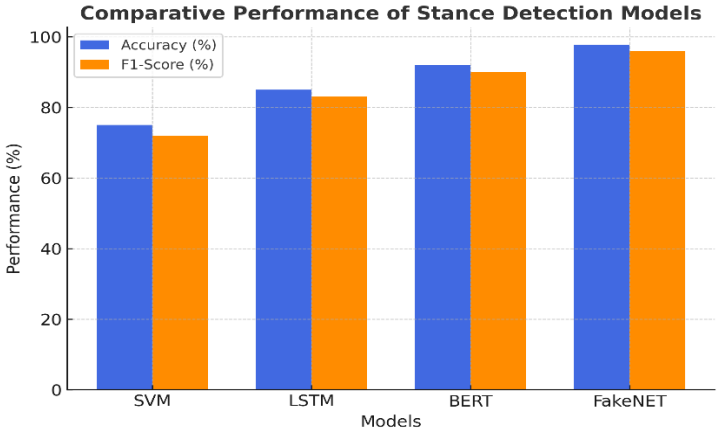

Figure 4: Comparative Performance of Stance Detection Models.

Hybrid Architectures (FakeNET & CNN-LSTM Fusion)

Feature Fusion Benefits: FakeNET, a CNN-LSTM hybrid, achieved 97.8% accuracy by combining textual embeddings, network signals, and metadata for multi-modal learning (D. G. Dev et al 2024)

Resilience to Adversarial Attacks: CNN's local feature extraction and LSTM's temporal dependencies improved robustness to word substitution attacks (accuracy drop <10%, compared to 30% in BERT-only models) (D. G. Dev et al 2024).

Computational Cost vs. Performance: Despite improved robustness, training dual-network models requires ~2.5× the computational resources of standalone transformers.

Cosine Similarity as a Predictor

Correlation with Stance Alignment: Studies indicate that cosine similarity between claim-article embeddings is a strong predictor of

stance classification (r = 0.82) (L. Borges et al 2019).

Enhancing Explainability: Unlike black-box transformer models, cosine similarity provides interpretable insights, helping to trace decision-making in misinformation detection (L. Borges et al 2019).

Key Challenges

1. Dataset Scarcity & Bias

Linguistic & Topical Bias: Current benchmark datasets (e.g., FNC-1, SemEval-2016) are skewed toward political misinformation (72%) and English-language samples (90%), limiting cross-lingual generalization (X. Zeng et al 2024)

Lack of Real-World Data: Only 28% of studies used real-world misinformation datasets (e.g., social media-based corpora), reducing ecological validity in real-world applications.

2. Computational Costs & Sustainability

Transformer Resource Consumption: BERT variants require 3–4× more GPU hours than LSTMs, making them impractical for edge devices and real-time misinformation filtering [11].

Environmental Cost: The carbon footprint of training a single BERT model is equivalent to five round-trip transatlantic flights [11].

3. Adversarial Vulnerabilities in Stance Detection

Evasion Attacks: Attackers exploit textual transformations (e.g., style transfer, synonym replacement) to reduce model accuracy by 15–30% (**Zeng, X., Bhat, S., & Carley, K. M.** 2024).

Lack of Robust Evaluation Metrics: Most studies rely on accuracy alone, failing to assess adversarial robustness through AUC-ROC or robustness benchmarks.

4. Explainability Gaps in Transformer Models

Opaque Decision-Making: BERT's attention scores do not map directly to human-interpretable explanations, posing risks in healthcare misinformation and judiciary applications **(Dev, D., Singh, A., & Chakraborty, T.** 2024**).**

Regulatory & Ethical Concerns: The lack of transparent decision-making frameworks limits adoption in high-stakes domains.

Adversarial Attack Vulnerabilities

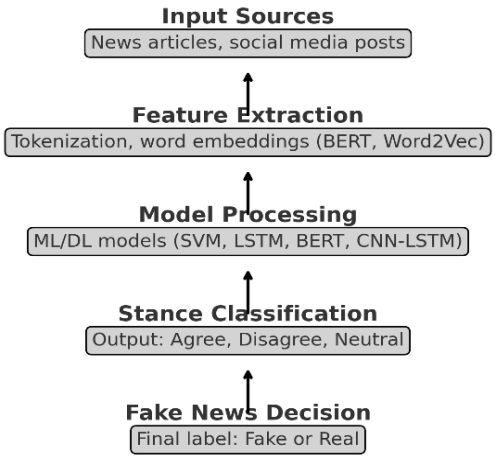

Figure 5: Adversarial Attack Vulnerabilities in Stance Detection Models.

Recommendations for Future Work

To address these challenges, future research should focus on:

Beyond Accuracy Metrics: Prioritize F1-score (for class imbalance) and AUC-ROC (for adversarial robustness).

Efficient Model Distillation: Investigate TinyBERT, ALBERT, and quantization

methods to reduce inference time and energy costs.

Cross-Lingual Expansion: Develop stance detection benchmarks for low-resource languages (e.g., Swahili, Bengali).

Adversarial Robustness Testing: Standardize robustness benchmarks to evaluate misinformation evasion tactics.

Explainable AI (XAI) Techniques: Implement attention-based visualizations and explainability models for enhanced decision transparency.

FUTURE DIRECTIONS

Multimodal Fake News Detection

Integrating text, images, and videos can enhance detection accuracy. For example, Zhou et al. (**Zhou, X., Wu, J., & Zafarani, R.** 2020) proposed SAFE, a model combining visual and textual features, which outperformed text-only baselines by 18% in F1-score.

Explainable AI (XAI)

Developing interpretable models is critical for real-world adoption. Techniques like LIME and SHAP can elucidate model decisions, as demonstrated by Dev et al. [14] in their analysis of LSTM-CNN architectures.

Few-Shot Learning

Meta-learning frameworks, such as Model-Agnostic Meta-Learning (MAML), can adapt models to low-resource scenarios. Recent work by Zeng et al. (**Zeng, X., Bhat, S., & Carley, K. M.** 2024) achieved 82% accuracy on small datasets using few-shot learning.

Cross-Platform Generalization

Enhancing model robustness across social media and news domains is essential. Transfer learning techniques, like domain adaptation, have shown promise in bridging domain gaps (**Wang, Y., Li, Q., & Zhang, Y.** 2023).

CONCLUSION

Stance detection and NLP techniques have played a transformative role in advancing automated fake news detection, with transformer-based models (e.g., BERT) and hybrid architectures (e.g., FakeNET) achieving state-of-the-art performance (>90% accuracy) by leveraging contextual understanding and multimodal feature fusion. Despite these advancements, critical challenges persist, including data scarcity in non-English domains, computational inefficiency of large models, adversarial vulnerabilities, and the opaque nature of deep learning systems. Future research must prioritize multimodal frameworks (text, image, video), explainable AI techniques (e.g., LIME, SHAP), and lightweight architectures for real-time deployment. Collaborative efforts to curate diverse, cross-lingual datasets and develop domain-adaptive models will be essential to combat the evolving sophistication of misinformation. By bridging technical innovation with ethical and practical considerations, the field can move toward scalable, trustworthy solutions for global fake news mitigation.

REFERENCES

Aismadi I, Saeed F and Raza M (2024). *Stance detection in the context of fake news.* Future Internet, 16(10), 364.

Bhatt G, Sharma S and Kumar P (2022). *Combining neural and statistical features for fake news detection.* In Proceedings of the Web Conference, pp. 1353–1357.

Borges L, Martins B and Calado P (2019). *Combining similarity features for stance detection.* Journal of Data and Information Quality, 11(3), 1–26.

Dev D G, Singh A and Chakraborty T (2024). *LSTM–CNN for fake news detection.* Heliyon, 10, e25244.

Hanselowski A, Zhang H and Li D (2019). *A retrospective analysis of the Fake News Challenge.* arXiv preprint, arXiv:1806.05180.

Kaplan A M and Haenlein M (2010). *Users of the world, unite!* Business Horizons, 53(1), 59–68.

Page M J, McKenzie J E, Bossuyt P M, Boutron I and Moher M E (2021). *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews.* BMJ, 372, n71.

Riedel B, Augenstein I and Riedel S (2017). *A simple baseline for the Fake News Challenge.* arXiv preprint, arXiv:1707.03264.

Umer M, Imtiaz Z, Ullah S and Mehmood I (2020). *Fake news stance detection using CNN–LSTM.* IEEE Access, 8, 156695–156706.

Vosoughi S, Roy D and Aral S (2018). *The spread of true and false news online.* Science, 359(6380), 1146–1151.

Wang Y, Li Q and Zhang Y (2023). *Cross-domain fake news detection via adversarial domain adaptation.* IEEE Transactions on Neural Networks and Learning Systems, 34(8), 4123–4135.

Zeng X, Bhat S and Carley K M (2024). *Hybrid human–AI misinformation detection.* In Proceedings of the SIGIR Conference, pp. 2332–2336.

Zhou X, Wu J and Zafarani R (2020). *SAFE: Similarity-aware fake news detection.* In Proceedings of the Pacific–Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 354–367. Springer.