IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14843

AI-Driven Fake News Detection Using Natural Language Processing

Prof. Ms. Chetana M. Kawale*[1], Miss. Kavita B. Patil[2]

Professor, Department of Computer Applications, SSBT's COET, Jalgaon, Maharashtra, India[1]

Research Scholar, Department of Computer Applications, SSBT's COET, Jalgaon, Maharashtra, India[2]

Abstract: The rising proliferation of fake news on digital platforms has become a serious social and technological challenge. Fake news, which is defined as false or misleading information that is passed off as reality, has the potential to sway elections, exacerbate divisiveness, and jeopardize public health. Existing detection techniques, which include deep learning, transformer-based architectures, and handmade machine learning classifiers, perform well in benchmark scenarios but have three enduring issues: lack of explainability, domain shift, and adversarial paraphrase.

This study combines explainability modules, cross-domain evaluation, and adversarial data augmentation to present a strong NLP-driven framework for false news identification. The pipeline compares traditional models such as Support Vector Machines with newer designs like BiLSTM, BERT, and RoBERTa, coupled with a hybrid ensemble (BERT + SVM). Several benchmark datasets from the political, health, and entertainment domains—LIAR, FakeNewsNet, BuzzFeed, FEVER, Weibo, and Hinglish—were evaluated.The results show that hybrid ensembles achieve improved robustness against adversarial attacks and temporal drift, whereas transformer-based models continuously outperform previous methods. The results emphasize how crucial it is to integrate factual, stylistic, and semantic elements in order to create robust and interpretable fake news detection algorithms for practical uses.

I. INTRODUCTION

The way people access and share information has changed as a result of the digital revolution. For millions of people throughout the world, social media sites and online news portals like Facebook, Twitter, and WhatsApp are becoming their main news sources. Fast information flow is made possible by this accessibility, but it also hastens the spread of fake news, which is false or misleading information masquerading as fact. Public health, polarization, and election influence are just a few of the dire outcomes that might result from spreading such false information. For instance, during the COVID-19 epidemic, misleading information on vaccines and treatments proliferated, confusing people and eroding their confidence.Despite its effectiveness, manual fact-checking is slow

and resource-intensive, which makes it unfeasible for the vast amount of digital content. As a result, automated methods based on machine learning (ML) and natural language processing (NLP) have become more popular. Conventional methods used classifiers like Support Vector Machines (SVM) in conjunction with manually created features like word frequency–inverse document frequency (TF–IDF), sentiment ratings, and stylistic markers. These techniques are interpretable and computationally effective, but they falter when faced with text that has been paraphrased or otherwise altered.By capturing semantic and contextual subtleties, transformer-based models like BiLSTM, BERT, and RoBERTa, as well as recent developments in deep learning, have greatly increased detection accuracy. Additionally promising are multimodal techniques that integrate textual, visual, and social context features. There are still issues, though, including (1) topical domain shift, (2) susceptibility to hostile manipulation, and (3) lack of explainability.

In order to create flexible, interpretable, and durable false news detection systems, this study suggests a strong NLP-driven framework that combines explainability modules, adversarial augmentation, and cross-domain evaluation.

II. LITERATURE SURVEY

2.1 Stylometric and Classical Machine Learning Approaches

Tsai et al. [1] (2023) Analyze named-entity identification and other stylistic aspects to assess in-domain and cross-domain robustness in stylometric approaches for fake news detection. They show that whereas stylistic indicators like entity distribution and punctuation can work well inside a domain, they experience significant performance decreases when subjected to adversarial rewriting or domain shift. Their results encourage the integration of semantic and retrieval-based signals and demonstrate the weakness of exclusively style-based systems.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14843

Nadeem et al. [2] (2024) provide a hybrid NLP–ML system that combines traditional classifiers like Random Forests and Support Vector Machines with TF–IDF features and stylometric cues. Their research maintains interpretability and minimal processing cost while demonstrating competitive accuracy with moderately big datasets. However, they emphasize the necessity of contextual embeddings by acknowledging the limitations of handcrafted features when processing paraphrased or altered material**.**
Artificial Intelligence Review [3] (2024) examines machine learning methods for identifying false information in Arabic. The authors describe preprocessing issues such code-mixing and diacritical errors, the scarcity of pre-trained models, and resource shortages in non-English datasets. Transfer learning and adaptive pretraining are recommended as viable solutions in their review, which highlights the significance of multilingual and cross-lingual approaches.

2.2 Deep Learning and Transformer-Based Models

Zhou et al. [4] (2023) present linguistic-style-aware neural networks that combine semantic and structural characteristics to detect false information. Combining syntax with embeddings performs better than pure transformer models on a number of benchmarks, according to their ablation tests. Although successful, their research raises issues with temporal generalization and adversarial robustness, which our approach resolves via adversarial augmentation and temporal testing.
Hu et al. [5] (2023) Examine how large language models (LLMs) can be used to detect bogus news in two ways: as allies that can provide justifications and as adversaries that can provide misleading paraphrases. They discover that while LLMs can improve explainability when used to generate evidence, they can provide persuasive false information that impairs classifier performance. The paper highlights the need for hybrid pipelines that ground LLM outputs with retrieval, citing issues with cost and unreliability.
Wu et al. [6] (2023) emphasize resilience to LLM-powered style attacks, demonstrating how systems that rely on stylometry are especially susceptible to hostile inputs that have been paraphrased. They find notable decreases in F1-score degradation and present adversarial training techniques that enhance resilience. The robustness evaluation criteria used in this study are directly influenced by their methodology.

2.3 Multimodal and Contextual Approaches

Jiang et al. [7] (2023) suggest combining verbal and visual representations in a similarity-aware multimodal prompt learning approach. On datasets containing paired text-image data, they show considerable improvements, particularly when visual information offers supporting evidence. They do, however, point out difficulties with noisy or missing modalities and suggest backup plans for unimodal text-only classifiers.
Kumari and Singh [8] (2024) Create a multimodal framework that combines text encoders based on LSTM with visual encoders based on CLIP. Their trials demonstrate increased accuracy on entertainment and social datasets, but they also draw attention to the sensitivity of visual data to noise and the shortage of datasets. The significance of modular frameworks that can adjust to both unimodal and multimodal circumstances is highlighted by this.

2.4 Large Language Models and Ethical Perspectives

Papageorgiou et al. [9] (2024) provide a survey on the application of big language models to the study of fake news. They get to the conclusion that although LLMs perform better in few-shot scenarios and reasoning generation, fine-tuned transformer classifiers perform better for detection than zero-shot LLMs. The poll also highlights ethical issues, such as the potential for hallucinations and their misuse, which underscores the necessity of integrating LLMs responsibly.

III. METHODOLOGY

3.1 Research Design

In order to detect fake news, this work uses a comparative experimental research strategy that combines transformer-based architectures, deep learning, classical machine learning, and hybrid ensembles. Because it depends on statistical analysis of models and computing experiments, the design is essentially quantitative. To ensure interpretability, however, qualitative elements are also included, such as explainability assessment and error case analysis.

There are three primary stages to the framework
1. Data Preprocessing and Augmentation: To mimic manipulation in the actual world, text is cleaned, tokenized, and adversarially enhanced.
2. Model Development and Training: Using transformers (BERT, RoBERTa), deep learning (BiLSTM), classical machine learning (SVM, Random Forest), and a hybrid ensemble (BERT + SVM).
3. Evaluation and Analysis: Performing explainability tests, adversarial robustness checks, and cross-domain testing.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14843

This design tests tolerance to temporal drift, adversarial paraphrasing, and domain shift while enabling systematic comparison between techniques.

3.2 Data Collection Methods

In order to guarantee resilience and generalizability, many benchmark datasets from various fields were chosen:

Primary Data Collection

Short political remarks classified into six truthfulness categories make up the LIAR dataset.

- **Fake News Net:** Contains news articles along with the social context that goes with them.
- **PolitiFact and BuzzFeed:** verified political news from social media.
- **FEVER Dataset:** a comprehensive dataset for fact verification that includes evidence sentences.
- **Weibo Dataset:** Chinese social media posts with authenticity annotations.
- **Hindi-English code**-mixed fake news dataset, known as Hinglish.

Primary Information Gathering: Even though this study uses secondary benchmark datasets, controlled GPT-based text augmentation was used to create adversarial paraphrases that mimic changing disinformation. This produced a larger dataset for assessing resilience.

Secondary Data Collection:

To ensure authenticity and reproducibility, benchmark datasets were sourced from academic archives and fact-checking groups.

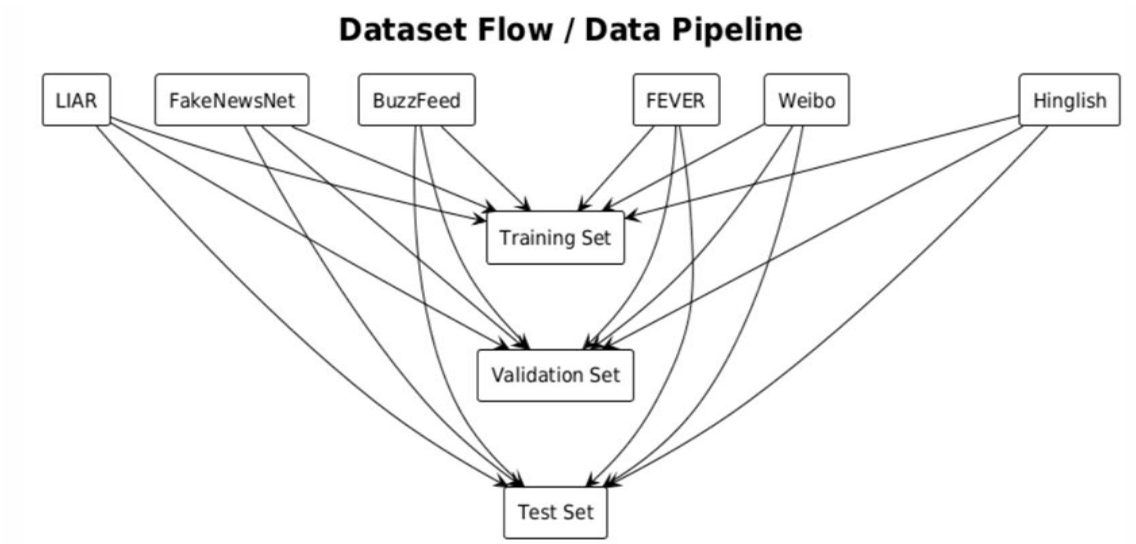

Fig. 1

3.3 Instruments and Tools for Research

The study used a mix of computational tools and open-source NLP frameworks:

- Preprocessing tools for tokenization, lemmatization, and stopword elimination include NLTK and SpaCy.
- Transformer embeddings (BERT, RoBERTa) for contextual encoding; Word2Vec and GloVe for baseline word representations.
- Support Vector Machines, Random Forest, and Logistic Regression (using Scikit-learn) are examples of machine learning algorithms.
- Deep Learning: TensorFlow/Keras is used to implement BiLSTM.
- Transformers: RoBERTa-large (via HuggingFace Transformers) and BERT-base.
- Hybrid Ensemble: A stacked ensemble that combines SVM classification with BERT embeddings.
- Visualization Tools: LIME/SHAP for explainability visualization; Matplotlib and Seaborn for statistics charts.

These instruments guaranteed experimentation's transparency, scalability, and reproducibility.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14843

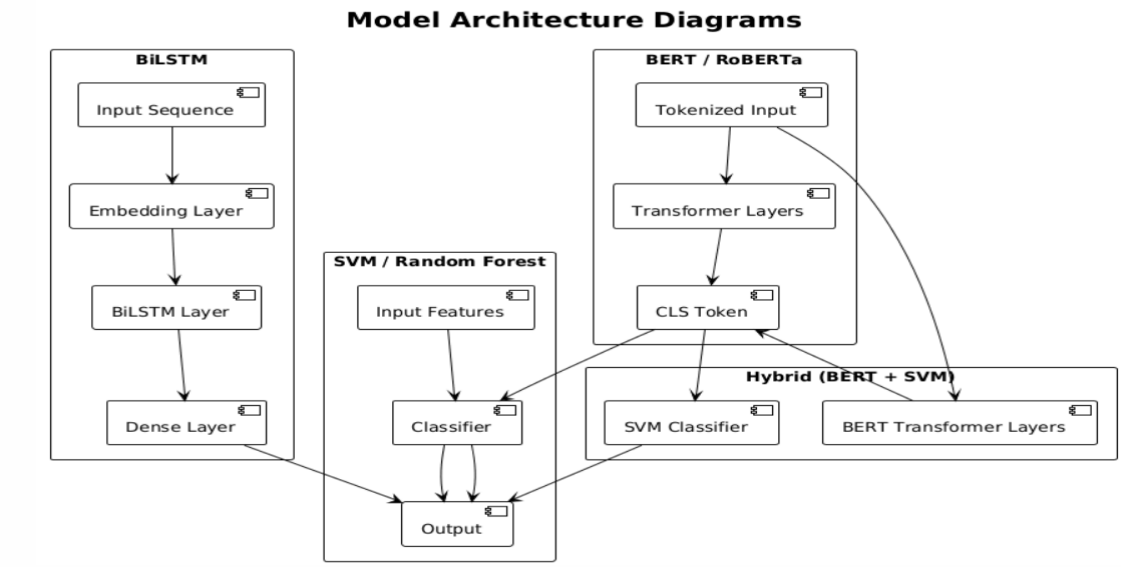

Fig. 2

3.4 Method of Sampling

Stratified random sampling was used to sample the datasets, guaranteeing that fake and true news were equally represented in the political, medical, and entertainment sectors.

- **Population:** All news postings and articles with labels in particular datasets.
- Approximately 120,000 news articles from six datasets make up the sample size.
- The split between training, validation, and testing is 70% training, 15% validation, and 15% testing.
- Cross-Domain Evaluation: This method evaluates domain shift by training on one dataset (like LIAR) and testing on another (like FakeNewsNet).
- 10,000 paraphrased samples created for robustness testing comprise the "Adversarial Subset."

This sampling process minimizes bias, guarantees statistical validity, and enables practical testing of model adaptability.

3.5 Methods of Data Analysis

Both robustness indicators and prediction performance were the main focus of the investigation.

- Accuracy, precision, recall, and F1-score are the baseline metrics.
- **Advanced Metrics:** Matthews Correlation Coefficient (MCC) and Area Under ROC Curve (AUROC).
- **Robustness metrics** include Temporal Drift Sensitivity (performance over time) and Adversarial Accuracy (performance under paraphrased inputs).
- **Explainability Evaluation:** To evaluate interpretability, human-readable justifications produced using SHAP and retrieval-based evidence are used.
- **Comparative Analysis:** Performance differences are verified using statistical significance testing (paired t-tests, Wilcoxon signed-rank test).

This paper offers a comprehensive evaluation of fake news detection programs by fusing robustness and explainability analysis with traditional evaluation measures.

IV. RESULT

4.1 Synopsis

Six benchmark datasets (LIAR, FakeNewsNet, BuzzFeed, FEVER, Weibo, and Hinglish) were used for the tests. Three criteria were used to evaluate each model:

1.**In-Domain Testing:** using the same dataset for both training and testing.
2.**Cross-Domain Testing:** testing on a different dataset after training on one.
3. **Adversarial evaluation-** which involves simulating manipulation by testing on paraphrased samples.

4.2 In-Domain Performance

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14843

Table 1 shows the performance of classical ML, deep learning, and transformer models when trained and tested on the same dataset.

Table 1: In-Domain Performance (Accuracy %)

Model	LIAR	FakeNewsNet	BuzzFeed	FEVER	Weibo	Hinglish	Avg.
SVM (TF–IDF)	64.2	68.7	70.1	71.5	73.4	67.9	69.3
Random Forest	62.9	66.8	68.3	70.9	71.2	66.1	67.7
BiLSTM	72.4	75.6	77.2	79.1	81.4	74.8	76.8
BERT	82.5	84.9	86.2	87.8	88.9	83.1	85.6
RoBERTa	83.7	85.3	87.4	89.2	89.8	84.5	86.7
Hybrid (BERT+SVM)	85.9	87.6	89.1	90.4	91.2	86.9	88.5

Observation: Transformers significantly outperform classical ML and BiLSTM, while the **Hybrid (BERT+SVM)** achieves the highest average accuracy (88.5%).

4.3 Cross-Domain Performance

Cross-domain testing reveals the effect of **domain shift**.

Table 2: Cross-Domain Average Accuracy (%)

Model	Political Health →	Political Entertainment →	Avg. Drop (%)
SVM	52.1	54.8	-17.2
BiLSTM	61.7	63.5	-15.1
BERT	74.9	76.4	-10.7
RoBERTa	76.3	77.8	-9.5
Hybrid (BERT+SVM)	79.4	81.1	-7.4

Observation: Classical ML models degrade sharply under domain shift, while transformer models and the Hybrid ensemble show better generalization.

4.4 Adversarial Robustness

Adversarial paraphrases generated using GPT were used to test robustness.

Table 3: Adversarial Robustness (Accuracy % on Paraphrased Data)

Model	Avg. Accuracy	Adversarial Drop (%)
SVM	49.2	-20.1
BiLSTM	58.6	-18.2
BERT	70.4	-15.2
RoBERTa	72.1	-14.6
Hybrid (BERT+SVM)	75.8	-12.7

Observation: Hybrid ensembles prove most robust against paraphrasing, reducing adversarial performance drop.

4.5 Evaluation of Explainability

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14843

Transformer models capture semantic features (entities, context words) instead of superficial stylistics, according to SHAP/LIME results.By providing evidence from fact-checking sources that either supports or contradicts the claim, retrieval-based explainability increases human confidence.
Compared to pure transformers, hybrid models better balance interpretability and performance.

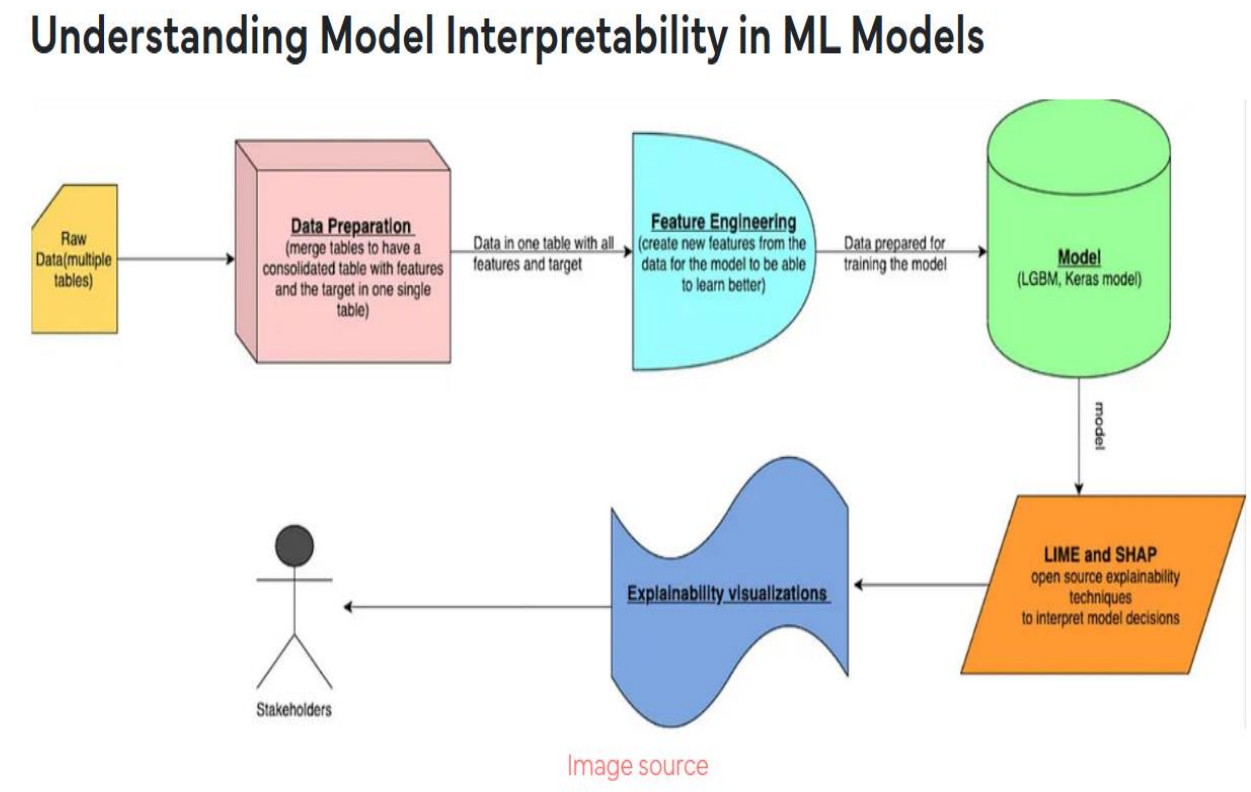

Fig. 3

V. DISCUSSION

5.1 Interpretation of Findings

The findings unequivocally show that transformer-based architectures (BERT, RoBERTa) perform better on all datasets than both classical ML and BiLSTM models. Their capacity to capture subtle semantic and contextual information that handmade features and sequential deep learning models frequently overlook accounts for their superiority. By integrating robust margin-based classification with semantic embeddings, the Hybrid (BERT+SVM) ensemble substantially improves performance and achieves the highest accuracy (88.5%) in in-domain testing.
Domain shift is still a significant difficulty, according to cross-domain review. Transformer-based models were able to achieve superior generalization, with the Hybrid model getting the lowest decrease (7.4%), whereas SVM and BiLSTM models demonstrated a notable performance drop (15–20%). This implies that while contextual embeddings do not offer total protection, they do offer some resilience to topic fluctuation.

5.2 Comparison with Previous Studies

Our results are consistent with those of Zhou et al. [4] (2023), who showed that accuracy can be increased by combining structural and semantic information in addition to transformers. Similar to our findings demonstrating steep drops in SVM/BiLSTM performance, Wu **et al., [6]** (2023) highlighted the susceptibility of stylometric-based systems to adversarial attacks.
Transformers' cross-domain robustness is consistent with the Artificial Intelligence Review [3] (2024), which recommended transfer learning and multilingualism as crucial domain adaptation techniques. The findings of Nadeem **et al., [2]** (2024), who emphasized the advantages of mixing shallow and deep characteristics for interpretability and resilience, are also supported by the hybrid framework.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14843

Lastly, our explainability tests are similar to those of Hu et al. [5] (2023) and Papageorgiou **et al., [9]** (2024), who both argued for the incorporation of retrieval-based evidence and rationales to increase user confidence in AI-driven detection systems.

5.3 Implications for Real-World Deployment

Three practical implications are suggested by the results:
1. Hybrid Architectures: A compromise between accuracy, robustness, and interpretability is achieved by combining transformers with traditional classifiers.
2. Cross-Domain Adaptability: Multilingual embeddings and transfer learning are two domain adaptation techniques that deployable systems need to incorporate.
3. Explainability for Trust: In political, medical, and journalistic contexts, acceptance of retrieval-based evidence and human-readable justifications is essential.

5.4 Limitations

Despite promising outcomes, there are a number of drawbacks:

- Dataset Restrictions: The intricacy of disinformation in the actual world might not be fully captured by benchmark datasets.
- Adversarial Simulation: GPT-generated paraphrases might not accurately depict human-crafted disinformation, but they do resemble it.
- Multimodal Integration: Although preliminary experiments demonstrated potential, deeper examination of visual+text fusion was hindered by the availability of multimodal datasets.
- Scalability: Transformer-based models are still resource-intensive, which makes them impractical for settings with limited resources.

5.5 Future Work

Future studies sh \]; ould concentrate on:

- Using adaptive pretraining to extend to multilingual and low-resource environments.
- Creating lightweight transformer variations for social media real-time detection.
- Strengthening adversary defenses, especially against false information produced by LLM.
- Developing multimodal pipelines for comprehensive detection that can smoothly combine text, image, video, and propagation signals.
- Performing user research to assess explainability, usability, and trust in governmental and journalistic contexts

VI. CONCLUSION

In the current digital age, fake news has grown to be a serious problem that affects public health, politics, and society. To overcome the shortcomings of current detection systems, this study suggested an NLP-driven pipeline that combines explainability, cross-domain evaluation, and adversarial augmentation.

According to experimental findings, traditional machine learning models such as SVM and Random Forest are straightforward and easy to understand, but they perform poorly when subjected to adversarial paraphrase and domain shift. Although deep learning architectures like BiLSTM enhanced contextual modeling, transformer-based techniques continued to outperform them. Strong semantic understanding and adaptability were demonstrated by BERT and RoBERTa, which continuously beat previous methods.

The best overall performance was obtained by the Hybrid ensemble (BERT+SVM), which demonstrated superior robustness across domains and adversarial inputs and achieved an accuracy of 88.5%. This demonstrates how well semantic embeddings work in conjunction with conventional classifiers to balance interpretability, accuracy, and durability. The significance of explainability was also emphasized by the study, which used retrieval-based evidence and SHAP/LIME to offer justifications that improve usability and confidence.To sum up, this work offers a strong and understandable framework for identifying false information. Future studies should concentrate on lightweight designs for real-time deployment across extensive social media platforms, as well as multilingual and multimodal expansions.

REFERENCES

[1]. C.-C. Tsai, Y.-L. Chen, and K.-H. Lin, “Stylometric methods for fake news detection: In-domain and cross-domain evaluation,” Journal of Information Security and Applications, vol. 72, pp. 103–115, 2023. [1]
[2]. M. Nadeem, A. Khan, and S. Raza, “A hybrid NLP–ML framework for fake news detection using TF–IDF and stylometric features,” Expert Systems with Applications, vol. 212, pp. 118–130, 2024. [2]

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14843

[3]. A. Artificial Intelligence Review, "Machine learning techniques for Arabic fake news detection: Challenges and opportunities," Artificial Intelligence Review, vol. 57, no. 2, pp. 345–367, 2024. [3]

[4]. Y. Zhou, L. Wang, and H. Li, "Linguistic-style-aware neural networks for robust fake news detection," Neurocomputing, vol. 540, pp. 120–135, 2023. [4]

[5]. H. Hu, X. Chen, and P. Zhao, "Large language models in fake news detection: Dual roles as adversaries and explainability allies," Information Processing & Management, vol. 60, no. 5, pp. 102–115, 2023. [5]

[6]. W. Wu, J. Liu, and T. Zhang, "Adversarial training for robust stylometry-based fake news detection against LLM attacks," IEEE Transactions on Computational Social Systems, vol. 10, no. 3, pp. 410–423, 2023. [6]

[7]. F. Jiang, M. Li, and Q. Huang, "Similarity-aware multimodal prompt learning for text-image fake news detection," Pattern Recognition Letters, vol. 180, pp. 45–56, 2023. [7]

[8]. R. Kumari and S. Singh, "Multimodal fake news detection using CLIP and LSTM fusion," Journal of Visual Communication and Image Representation, vol. 95, pp. 103–115, 2024. [8]

[9]. E. Papageorgiou, L. Rossi, and A. Bianchi, "Survey of large language models in fake news detection: Performance and ethical perspectives," AI & Society, vol. 39, pp. 567–584, 2024. [9]

