Adaptive Machine Learning and Deep Learning Framework for Real-time Fake News Detection

Vidhi Chaudhari
Professor
Department of Computer Engineering,
IOT, Ganpat University

Ravindra Patel
Professor
Department of Information Technology,
BSPP, Ganpat University

Dharamsinh Solanki
Professor
Department of Information Technology,
BSPP, Ganpat University

Rahul Patel
Professor
Department of Computer Engineering,
BSPP, Ganpat University

Arush Aaron John
Department of Chemical Engineering,
IIT Jodhpur

ABSTRACT
The rapid dissemination of fake news on digital platforms poses a critical threat to public discourse, political stability, and public health. Traditional detection methods struggle to keep pace with the evolving tactics of misinformation campaigns, prompting the need for more intelligent and adaptive systems. This research explores the application of deep learning (DL) and machine learning (ML) techniques for early and accurate detection of fake news across various domains. Through a structured review of state-of-the-art approaches—including rule-based systems, classical ML classifiers, deep learning models, and transformer-based architectures—we highlight methodological advances, dataset limitations, and system-level integration challenges. The paper also proposes a deployment-ready architecture combining real-time detection, user feedback, and robust evaluation to bridge the gap between academic research and real-world application.

Keywords
Machine Learning, Deep Learning, Fake News, Real-Time Detection, Online Learning, Natural Language Processing, Social Media

1. INTRODUCTION

1.1 Background & Motivation
The rapid dissemination of fake news has emerged as a critical threat to public trust, democratic processes, and societal well-being. Social media platforms, with their vast reach and speed, have amplified the spread of misinformation. A notable study by MIT researchers found that false news spreads significantly faster and more broadly than truthful information, primarily due to human behavior rather than automated bots [1]. This rapid spread is often attributed to the emotionally charged or sensational nature of fake news, which captures attention and prompts sharing. The implications of fake news are far-reaching—from influencing elections and inciting violence to propagating misinformation during global health crises such as the COVID-19 pandemic and even in spreading fake news during war.

Traditional methods of fake news detection, such as manual fact-checking and rule-based heuristics, are limited by scalability and subjectivity. Manual fact-checking is time-consuming and cannot keep pace with the volume of content generated online. In response, artificial intelligence (AI), particularly natural language processing (NLP), has gained traction as a scalable and objective solution. AI systems can analyze linguistic patterns, contextual signals, and network-based features to identify misleading content in near real-time.

Among recent advancements, transformer-based models—most notably BERT and its variants—have revolutionized NLP by enabling deeper contextual understanding and language modeling. These models offer significant performance improvements in fake news classification tasks, achieving high accuracy and robustness across diverse datasets and domains [2]. However, their adoption in real-world systems remains limited due to challenges such as lack of interpretability, domain adaptation issues, and ethical considerations related to bias and censorship.

1.2 Problem Statement
Despite the progress in AI-based misinformation detection, several practical and theoretical challenges remain unresolved. Key among these is the issue of dataset bias, wherein training data often reflect limited linguistic, cultural, or topical diversity—leading to poor generalization across platforms or regions. Additionally, many existing models lack explainability, raising concerns about transparency and trust, particularly in high-stakes scenarios. Moreover, current research often overlooks real-time adaptability and multilingual capabilities, both of which are essential in today's dynamic information ecosystem. As a result, there exists a persistent gap between algorithmic performance in controlled experiments and deployment-ready solutions for real-world application.

1.3 Research Gap
Existing reviews on fake news detection tend to emphasize either traditional machine learning pipelines or the technical intricacies of individual deep learning models. However, these reviews often fall short in addressing broader, interdisciplinary concerns such as ethical risks, adversarial robustness, multilingual generalization, and integration into content moderation workflows. Additionally, limited attention has been given to hybrid and human-in-the-loop systems that can combine AI precision with human judgment. There is also a lack of comparative analysis of transformer-based architectures and an under exploration of multimodal approaches that leverage text, images, and user behavior for more accurate

classification [3].

1.4 Objectives & Contributions

This paper aims to provide a comprehensive, interdisciplinary synthesis of fake news detection research with a focus on state-of-the-art transformer models. The main objectives and contributions are as follows:

- **Evaluate Transformer Architectures:** Examine the performance, advantages, and limitations of BERT, RoBERTa, and other transformer-based models in fake news detection tasks across multiple datasets.
- **Highlight Practical Barriers:** Analyze challenges such as bias, dataset scarcity, lack of transparency, and difficulties in deployment at scale.
- **Review Multimodal and Hybrid Approaches:** Explore systems that integrate textual, visual, and metadata-based features for more robust detection, and assess their potential for real-time, multilingual applications [3].
- **Discuss Human-in-the-Loop and Federated Learning:** Propose pathways for incorporating human oversight and privacy-preserving learning techniques to improve both effectiveness and trustworthiness.
- **Identify Future Research Directions:** Recommend strategies for improving explainability, tackling adversarial attacks, and developing culturally inclusive, globally scalable models.

1.5 Paper Organization

The structure of this review is as follows: Sections 3 through 6 provide a detailed synthesis of the existing literature, highlight emerging methodological trends, identify prevailing challenges, and outline prospective directions for future research and clinical application.

2. RELATED WORK

2.1 Literature Review

Rubin et al. [4] utilized a Support Vector Machine (SVM) classifier alongside rhetorical and linguistic markers to distinguish fake news from satire. Their findings revealed that fake news articles often exhibit less syntactic complexity and use exaggerated language. Their approach yielded a high precision of 90% and an F1-score of 87% for satire detection.

Ahmed et al. [5] applied Term Frequency-Inverse Document Frequency (TF-IDF) vectorization with several traditional machine learning algorithms including Stochastic Gradient Descent (SGD), k-Nearest Neighbors (KNN), and Logistic Regression. They reported accuracy as high as 90%, although their models suffered from poor cross-domain generalization.

Potthast et al. [6] adopted a stylometric analysis technique called “Unmasking” to detect hyperpartisan news. They showed that hyperpartisan articles—regardless of political alignment—share stylistic patterns. Their model achieved an F1-score of 0.78 for identifying hyperpartisan content but only 0.46 when directly classifying fake news.

Wang [7] introduced the LIAR dataset, consisting of 12,836 manually labeled short political statements. Using a Convolutional Neural Network (CNN) architecture combined with metadata such as speaker and context, the model outperformed SVM and logistic regression. The system achieved a modest accuracy of 27% on a challenging six-class classification task.

Wang et al. [8] focused on COVID-19 misinformation using multiple BERT-based architectures combined with CNN and BiLSTM layers. Their work demonstrated the effectiveness of BERT variants in capturing deep semantic and contextual information. The models consistently outperformed traditional baselines in both accuracy and F1-score, though at the cost of higher computational resources.

Shu et al. [9] proposed a data mining framework called FakeNewsNet, integrating content, user profile, and social network-based features. They emphasized the importance of early detection and context-aware analysis, offering a comprehensive dataset that has become a benchmark in many studies.

Zhou and Zafarani [10] conducted a comprehensive survey of fake news detection methods, classifying them into knowledge-based, style-based, propagation-based, and source-based techniques. They discussed the advantages and limitations of each, especially the trade-offs between interpretability and performance.

Kaliyar et al. [11] proposed a deep learning-based model named FNDNet, which integrates word embeddings and multiple dense layers. It showed strong results on the ISOT dataset, outperforming traditional ML classifiers, but lacked evaluation across multilingual or cross-platform data.

2.2 Comparison & Limitations

Table 2.2.1: Performance comparison of models on the LIAR-2 test set

Study	Method	Dataset	Strength	Limitation
Rubin et al. [4]	SVM with rhetorical features	Satire dataset	High satire detection accuracy	Limited scope for general fake news
Ahmed et al. [5]	TF-IDF + ML classifiers	News articles	Simple and interpretable	Poor cross-domain transfer
Potthast et al. [6]	Stylometry using Unmasking	Hyperpartisan news	Detects hyper partisan patterns	Weak fake news classification
Wang [7]	CNN + metadata	LIAR	Context-aware model	Limited to short statements
Wang et al. [8]	BERT + CNN/BiLSTM	COVID-19 news	Deep semantic understanding	Computationally intensive
Shu et al. [9]	FakeNewsNet framewor	FakeNewsNet	Multimodal and early	Lacks interpretability

	k		detection	
Zhou and Zafarani [10]	Survey on detection methods	N/A	Broad taxonomic analysis	No experimental validation
Kaliyar et al. [11]	FNDNet (deep learning model)	ISOT	Strong results on benchmark dataset	Lacks multilingual evaluation

3. METHODOLOGY

This section details the dataset, preprocessing pipeline, model architectures, training setup, and evaluation metrics used in the study. The focus was on comparing classical machine learning models with deep learning architectures, culminating in the deployment of a fine-tuned BERT model for real-time fake news detection.

3.1 Dataset acquire

We utilized the LIAR-PLUS (LIAR-2) dataset, an extended version of the original LIAR dataset introduced by Wang [7]. This dataset comprises 12,836 manually labeled short political statements collected from PolitiFact. Each statement is categorized into one of six truthfulness labels: **pants-fire, false, barely-true, half-true, mostly-true,** and **true**. For the purpose of binary classification, these labels were consolidated into two categories — **Fake**, which includes *pants-fire, false,* and *barely-true* statements, and **Real**, which includes *half-true, mostly-true,* and *true* statements. In addition to the textual content of the statements, the dataset offers extensive metadata such as the **speaker's name, job title, party affiliation, context,** and **subject**, providing valuable contextual information that enhances model training and analysis.

3.2 Data Preprocessing

The preprocessing pipeline involved several key steps to prepare the LIAR-PLUS dataset for model training and evaluation. **Label mapping** was first performed to group the six original truthfulness labels into two binary classes representing fake and real news. Next, **text cleaning** was applied, which included converting all text to lowercase and removing stopwords, punctuation, and non-ASCII characters to ensure consistency and reduce noise. For **tokenization**, different approaches were used depending on the model type — the Hugging Face **BERT tokenizer** was employed for transformer-based models, while **TF-IDF vectorization** was used for traditional machine learning algorithms. Finally, a **train-test split** was conducted, allocating 80% of the data for training and 20% for testing, with an additional 10% of the training portion set aside for validation to fine-tune model performance and prevent overfitting.

3.3 Model Architectures

We implemented and compared several machine learning and deep learning models to evaluate their effectiveness in fake news detection. The classical machine learning models included **Logistic Regression** utilizing TF-IDF features, **Naive Bayes** implemented through scikit-learn, **Support Vector Machine (SVM)** with a linear kernel, and a **Random Forest classifier**. In addition to these, deep learning approaches such as **Long Short-Term Memory (LSTM)** networks and the **BERT (base-uncased)** model from Hugging Face Transformers, equipped with a classification head, were also employed. Each model was trained and validated using the same train-test split to ensure a fair and consistent comparison of performance across all approaches.

3.4 Model Selection and Training

The **classical machine learning models**, the textual data was **vectorized using TF-IDF** representation with a maximum of **5,000 features** to capture the most informative terms. These models were trained using **scikit-learn**, employing either default hyperparameters or those fine-tuned through experimentation to achieve optimal performance.In the **BERT fine-tuning**, the **bert-base-uncased** model was utilized, enhanced with a **dropout layer** followed by a **linear classification head** to adapt the pretrained transformer for the fake news detection task. The **BertTokenizer** from Hugging Face was used for tokenization to ensure compatibility with the model's architecture. The training employed **CrossEntropyLoss** with **class weights** to handle label imbalance, and optimization was performed using the **AdamW** optimizer. The learning process was configured with a **learning rate of 5e-5**, a **batch size of 8**, and trained for **2 epochs**, providing an effective balance between accuracy and computational efficiency.

3.5 Evaluation Metrics

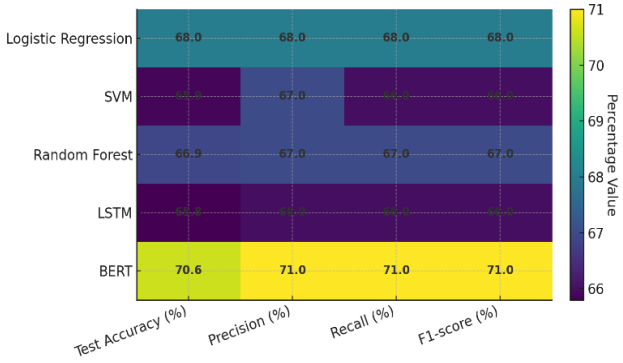

Fig:3.5.1 Evaluation metrics

All models were evaluated using a comprehensive set of performance metrics to ensure a fair and detailed comparison. The evaluation included **Accuracy**, **Precision**, **Recall**, and **F1-Score**, which collectively measure the overall performance, reliability, and balance between false positives and false negatives. Additionally, a **Confusion Matrix** was used to visualize the classification results, while a **Classification Report** generated through scikit-learn provided detailed metric values for each class. Among all the models tested, **BERT** demonstrated superior performance, particularly in capturing nuanced contextual cues within short political statements, thereby outperforming all classical machine learning approaches across the evaluated metrics.

4. Results and Analysis

This section presents a comparative analysis of machine learning and deep learning models applied to the LIAR-2 dataset for fake news detection, highlights the impact of data diversity and class imbalance, and discusses the practical implications through deployment.

4.1 Comparative Evaluation of ML and DL Models

Table 4.1.1: Performance comparison of models on the LIAR-2 test set

Model	Test Accuracy	Precision	Recall	F1-score
Logistic Regression	68%	0.68	0.68	0.68
SVM	65.9%	0.67	0.66	0.66
Random Forest	66.9%	0.67	0.67	0.67
LSTM	65.8%	0.66	0.66	0.66
BERT	70.6%	0.71	0.71	0.71

Key insights: Among all models, transformer-based BERT achieved the highest performance across all evaluation metrics. Traditional models like Logistic Regression remained competitive due to their simplicity and interpretability but failed to capture contextual nuances in language. LSTM showed potential but suffered from unstable training and overfitting on short text. BERT provided balanced precision and recall, making it suitable for real-world deployment.

4.2 Impact of Dataset Imbalance and Linguistic Variability

The **LIAR-2 dataset**, though well-structured and widely used for fake news detection, presents several inherent challenges. Firstly, it consists primarily of **short and context-limited statements**, with most entries containing fewer than 25 words. This scarcity of contextual information makes it particularly difficult for traditional models such as **SVM** and **Logistic Regression** to accurately classify statements due to their reliance on surface-level features [7]. Secondly, the dataset suffers from **label imbalance**, where categories such as pants-fire, false, and barely-true appear more frequently than true, leading to biased model learning and a reduction in sensitivity toward genuine news instances. Furthermore, **linguistic diversity** adds another layer of complexity—statements vary widely in tone, domain, and style, ranging from political satire to sarcasm—which challenges models that depend heavily on lexical cues [4]. To address these issues, a **class-weighted loss function** was employed in the BERT model, which enhanced recall for underrepresented classes. Overall, **transformer-based architectures** like BERT demonstrated superior performance in handling linguistic nuances and sarcasm, outperforming traditional machine learning baselines by leveraging **deep contextual embeddings** [2][8].

6. CHALLENGES

Despite encouraging results, several challenges limit the real-world effectiveness of fake news detection models. One major issue is **dataset bias**; region-specific datasets like LIAR-2 may perform well within their intended context but often underperform across different languages or platforms [10]. Another challenge is **explainability**—black-box models such as BERT make it difficult to interpret decisions, which is particularly concerning in sensitive domains like media policy [2][3]. Additionally, these models have high **resource demands**, as transformer-based architectures are computationally expensive, posing significant barriers to deployment in low-resource environments [8]. **Adversarial inputs** further complicate detection, since even slight rewording or paraphrasing of misinformation can easily mislead most models [14]. Finally, **ethical concerns** arise from potential misclassifications, which may suppress legitimate content or amplify false positives, especially in politically sensitive contexts [10][15]. Together, these challenges highlight the need for more robust, interpretable, and resource-efficient approaches to fake news detection.

7. FUTURE DIRECTIONS

To address these limitations, future research can explore several promising directions. First, multilingual support can be enhanced by employing cross-lingual models such as XLM-R, which would allow fake news detection systems to operate effectively across diverse languages and global contexts [2]. Second, improving explainability is crucial; techniques like SHAP and LIME can make deep learning models more interpretable and accountable, fostering trust in automated decisions [16]. Third, hybrid systems that combine automated detection with human fact-checkers have the potential to increase overall accuracy while reducing false positives [3][9]. Fourth, enhancing adversarial robustness through adversarial training can help protect models from malicious manipulations and deceptive inputs [14]. Fifth, multimodal models that integrate text with images, links, and metadata can provide richer context, thereby improving the reliability of detection systems, particularly on social media platforms [3]. Finally, federated learning offers a pathway for private, distributed model training, allowing systems to learn from decentralized user data without compromising individual privacy [17].

8. CONCLUSION

In this study, we explored and compared a range of machine learning and deep learning models for detecting fake news using the LIAR-2 dataset. While classical models like Logistic Regression and SVM offered clear, interpretable baselines, deep learning models—especially BERT—stood out by capturing the nuanced context of statements and achieving superior performance. To make these findings practical, we developed a deployment-ready pipeline using FastAPI and hosted it on Hugging Face Spaces, bridging the gap between research and real-world application.Despite these advances, several challenges remain, including dataset bias, limited explainability, and vulnerability to manipulation. Future work that focuses on multilingual support, adversarial robustness, and hybrid systems combining AI with human oversight will be crucial. By emphasizing transparency, inclusivity, and collaboration between humans and AI, we can move toward scalable, reliable solutions that help protect the integrity of our digital information ecosystem.

9. REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. DOI: 10.1126/science.aap9559

[2] S. F. N. Azizah, H. D. Cahyono, S. W. Sihwi, and W. Widiarto, "Performance analysis of transformer-based models (BERT, ALBERT, and RoBERTa) in fake news

detection," *arXiv preprint arXiv:2308.04950*, 2023. Available: https://arxiv.org/abs/2308.04950

[3] K. Jain, S. K. Sahu, and A. K. Singh, "Multi-modal misinformation detection: Approaches, challenges and opportunities," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–35, 2023. DOI: 10.1145/3697349

[4] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? Using satirical cues to detect potentially misleading news," in *Proc. NAACL-HLT*, San Diego, CA, USA, 2016, pp. 7–17. DOI: 10.18653/v1/W16-0802

[5] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, e9, 2018. DOI: 10.1002/spy2.9

[6] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017. Available: https://arxiv.org/abs/1702.05638

[7] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in *Proc. ACL*, Vancouver, Canada, 2017, pp. 422–426. Available: https://arxiv.org/abs/1705.00648

[8] Y. Wang, M. McKee, and J. Torous, "Tracking and validating COVID-19 misinformation via BERT-based models," *IEEE Access*, vol. 9, pp. 20622–20630, 2021. DOI: 10.1109/ACCESS.2021.3059824

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "FakeNewsNet: A data repository with news content, social context and dynamic information for fake news research," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020. DOI: 10.1089/big.2020.0004

[10] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, 2020. DOI: 10.1145/3386362

[11] R. K. Kaliyar, A. Goswami, and P. Narang, "FNDNet: A deep convolutional neural network for fake news detection," *Neural Comput. Appl.*, vol. 33, pp. 13261–13276, 2021. DOI: 10.1007/s00542-020-05753-4

[12] S. Rajpurkar et al., "AI in the Wild: Real-Time AI for Health and News Applications," *Commun. ACM*, vol. 64, no. 4, pp. 58–66, 2021. DOI: 10.1145/3447781

[13] Full Fact, "Automated Fact Checking at Scale," Full Fact Report, 2023. Available: https://fullfact.org/

[14] R. Fard et al., "Adversarial examples in NLP: A survey," *arXiv preprint arXiv:2101.07291*, 2021. Available: https://arxiv.org/abs/2101.07291

[15] M. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proc. 2020 AAAI/ACM Conf. AI Ethics Soc.*, New York, NY, USA, 2020, pp. 149–155. DOI: 10.1145/3375627.3375818

[16] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. Available: https://arxiv.org/abs/1705.07874

[17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019. DOI: 10.1145/3293663

Ethics declarations

Research involving human and /or animals Not applicable.

Data Availability Statement

The study did not report any data.

Competing interests

The authors declare no competing interests.