Misinformation Detection For Social Media News Using Transformer Model

Usman Ali
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
shanibhi964@gmail.com

Ahmad Hassan
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
mahmad91926@gmail.com

Sana Tariq
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
sana.tariq@iub.edu.pk

Iram Haider
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
iramhaider765@yahoo.com

***Abstract*—The advance of misinformation on digital platforms has appeared as a major social affair, influencing public opinion, destabilizing democratic processes, and undermining trust in institutions. Traditional detection methods—which rely on rule-based logic and keyword matching—have proven insufficient in addressing the emerging complexity of fraudulent content. Although deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have better performance, they continue to face limitations in scalability, generalization, and understandability. This paper presents a comprehensive framework for misinformation detection that constructs the use of the capabilities of transformer-based architectures, in particular BERT (Bidirectional Encoder Representations from Transformers). It increases with explainable artificial intelligence (XAI) techniques, including locally interpretable models (SHAP). These methods contribute to the understandability of model decisions, encourage transparency, and foster user trust. To further address the challenge of limited labeled datasets and domain-specific differences in misinformation, we include a semi-supervised learning policy. This integration enables the model to leverage both labeled and unlabeled data for better generalization across different disinformation domains. Using the political entity subset of the FakeNewsNet dataset, we conducted a comprehensive performance evaluation of our BERT-based structure in contrast to classical machine learning baselines such as Naive Bayes and Support Vector Machines (SVM). Experimental results show that our proposed model notably outperforms traditional methods, achieving high accuracy and providing understandable results. The framework demonstrates not only strong predictive power but also compliance and extensibility, key attributes for real-world deployment in fighting disinformation.**

***Index Terms*—Misinformation Detection, Transformer Models, Social Media, Fake News, BERT, Natural Language Processing, Deep Learning, Text Classification, News Verification**

I. Introduction

The rapid increase in the spread of misinformation on social media platforms has become a key social challenge, influencing public discourse, eroding the trust that is in institutions, and affecting democratic processes [1]–[4]. Customary detection Systems, such as rule-based classifiers or keyword filters, have struggled to keep pace with the increasing sophistication; there is content [5], [6]. These approaches often exhibit limited adaptability as well as semantic comprehension to detect certain context-aware misinformation patterns. In response, researchers have turned to ML and NLP for better ways to find deceptive content. [2], [7]. Deep learning models such as CNNs have been known to work to help prediction. Someday; however, they often suffer from high computational difficulty and bounded interpretability [8]–[11].

Lately, transformer-based models such as BERT have exhibited measurable improvements in language comprehension tasks [3]. These models can nearly catch long-range. The dependencies and the contextual nuances are in the text. This makes them highly effective for misinformation detection. However, while performance is promising, there remains a need for transparency in the decision-making processes, most of all in applications with high stakes such as news verification [12]. Explainable AI (XAI) techniques such as SHAP and LIME Offer solutions via providing interpretable justifications for the model's predictions [13], [14]. Integrating these techniques Using transformer models improves the trustworthiness and accountability of automated systems. the scarcity Labeled misinformation data motivates the use of semisupervised. Learning how to leverage both BH-labeled and unlabeled datasets [15], [16].

This paper puts forth a complete framework that combines a BERT-based architecture and explainable AI mechanisms and also deep learning strategies for tackling the problem of the detection of misinformation. We conduct many evaluations through the FakeNewsNet dataset [9]. Those results indicate further superior performance next to ML baselines that are customary, like NB plus SVM [13], [17].

II. Literature Review

Many of the checks have been on the issue of fake news. Detection via both customary and modern approaches. Initial

DOI: https://doi.org/10.67535/tsp.000001.015
Published under the ICACNC 2025 Proceedings

attempts centered on some handmade aspects, and those rule-based logics do include some lexical patterns. A further logic would be many sentiment scores. and user behavior signals no [7], [18]. While these methods Providing several basic understandings, they show some lack of strength in dynamic environments. With the dawning of deep learning, researchers began to use certain network models for fake news detection. For instance, worked to apply CNNs for extracting features that are hierarchical from the news articles [15]. Whereas they proposed a certain model of DNN–LSTM for detection from a URL [19]. Granting some effectiveness, those models face problems with transparency and computational cost. Transformer models have had some impact on NLP with a focus on attention. And skills with which to build dependencies [3]. BERT, in particular, has been widely adopted in much misinformation detection due to its improved context comprehension [10], [20]. BERT-based models greatly exceed standard ML methods across many datasets.

Deep learning models are, though, criticized often for being "black boxes." To address this, XAI techniques, such as SHAP. Values have been integrated to interpret model predictions and propose an interpretable model for trust in automated systems [17], [21]. Furthermore, they explored the use of explainability to assess bias and fairness in misinformation classifiers. To beat data scarcity, deep learning has gained traction [22]. Studies such as show that leveraging certain unlabeled data can greatly increase performance in a few low-resource scenarios applied [19]. Graph-based neural networks for a semi-supervised way to detect certain fake news in early times of propagation. Additionally, few researchers have incorporated multimodal data, including user profiles, propagation paths, and sources of credibility, to improve detection [9]. Also stressed the gains from weaving social settings into NLP pipelines. More recent work has focused on detecting deepfakes and also fake media, broadening misinformation's reach. To conclude, while current models provide sufficient foundations, the integration of transformers and explainable AI with semi-supervised learning presents a promising path ahead for strong and interpretable misinformation detection systems. In summary, while existing models provide strong foundations, the integration of transformers, explainable AI, and semi-supervised learning presents a promising path forward for robust and interpretable misinformation detection systems.

Despite quite large progress, some challenges remain: Data on limitations: Most models rely upon supervised learning, which requires large, annotated datasets that are expensive And time-consuming to prepare. Domain generalization: Models trained on just one data collection often fail to generalize across various domains, topics, or platforms. Model transparency: Highly accurate models like BERT prove difficult to interpret, which obstructs most adoption within sensitive applications. Real-time adaptation: Misinformation with tactics will evolve rapidly, requiring models that can dynamically adapt without having to be re-trained from scratch.

This paper addresses these gaps through a multidimensional framework or state-of-the-State performance in misinformation detection. The LIME and SHAP integrations are known to make model predictions explainable. Incorporates almost semi-supervised learning to use both labeled and unlabeled data, so it improves adaptability and generalizability. By combining these strategies, we present a solution that balances performance, understanding, and scalability. This is a major step towards a more dependable detection of misinformation systems.

To guide the reader, the paper is organized as follows:

- **Section I** introduces the background, motivation, and challenges in misinformation detection.
- **Section II:** Many customary approaches exist, including deep learning and transformer-based approaches.
- **Section III** tells of our method, as well as data processing. Classical baselines, BERT architecture, and also explainability tools and this deep learning.
- **Section IV** almost reports on the experimental results, comparing models and highlighting the effectiveness of XAI techniques.
- **Section V** concludes the study and discusses future work directions.

III. METHODOLOGY

The methodology of this study is designed to provide a comprehensive and scalable approach to disinformation detection by combining various components, each contributing to robustness, interpretation, and adaptation. The proposed framework follows a pipeline that includes data acquisition and preprocessing, baseline machine learning models, BERT-based transformer architecture, descriptive modules (LIME and SHAP), and semi-supervised learning augmentation

A. *Dataset*

We used the FakeNewsNet dataset, a well-known benchmark resource developed to aid research on fake news detection. Specifically, we focus on the PolitiFact subset, which contains fact-checked articles that have been classified as more or less real or fake based on expert journalistic analysis. The data collection includes the following attributes: News content: The full text of the article. Article title: A short headline that provides summary information. Label: A ground truth label (real or fake). Metadata (optional): Contains social context, source credibility, and publication timestamp. Dataset statistics Total articles: 1772 (combined fake and real categories) Language: English. Domains: Primarily political news. Distribution: Minor class imbalance (e.g., 973 fake, 779 real) pretreatment steps To prepare the data collection for modeling, we applied several standard NLP pretreatment techniques: Text cleaning: Removing HTML tags, URLs, and special

characters. Lower casing: normalizing case to reduce feature bulk. Stop word Removal: Removing common words (e.g., ”the,” ”is”) that add limited semantic value. Tokenization: Dividing text into tokens suitable for machine learning. Metadata linkage: Associating each article by its source, timestamp, and title where applicable. We also visualize a class distribution to verify minor imbalances. We then tailored our training strategy accordingly by adding class weights to our models.

B. *Dataset Preprocessing Models*

a) Naïve Bayes (multiple The input representation: each TF-IDF vector (word level, n-gram range: 1–2). Assumption: Features are mostly independent given the class. Advantages: Fast and interpretable Constraints: Unable to fully model dependencies in words. b) SVM Kernel: Linear (due to fast high-dimensional input) Input representation: TF-IDF vector Advantages: High accuracy on linearly separable data Limitations: Sensitive to noise; may overfit on sparse features. Each classifier was trained and validated using an 80/20 split. We used precision, accuracy, recall, and F1 scores as evaluation metrics.

C. *Transformer-Based Model (BERT)*

To take advantage of contextual understanding and capture important linguistic features, we implemented BERT (Bidirectional Encoder Representation from Transformers). We fine-tuned the pre-trained bert-base-uncased model from the HuggingFace Transformers library. Model Architecture Input: Article body sign using the WordPiece tokenizer. Embedding dimension: 768 Fine-tuning: Added a final classification layer on top of BERT’s [CLS] token output. Output: Binary classification (real vs. fake) Training order Loss function: Weighted cross-entropy (to remove class imbalance) Improviser: Adam W Batch size: 16 Learning rate: 2e-5 Epochs: 3-5 depending on initial stopping. We used GPU acceleration (via Google Colab) to reduce training time. The model achieved significantly higher performance than traditional models in terms of F1 score and accuracy.

D. *Deep Learning*

Since deep learning models are often criticized for being vague, we integrated explainable AI (XAI) tools: a) LIME (Local Interpretation Model-Agnostic Explanations) Process: Generates perturbed samples close to the prediction and trains a local interpretation model (e.g., linear regression). Usage: Describes how a certain article was deemed to be ”fake” via keywords and terms. Advantage: Completely model-agnostic and also readable. Process: Calculates certain Shapley values based upon game theory to feature contributions. Usage: Global explanations exist for the general overall behavior of that model, as local explanations by each instance. Advantage: It provides users consistency in theoretical guarantees of contribution values. We used several tools for creating explanations that are clear for predictions that enable us to identify all words influencing decisions from the model. This fully improves overall transparency, builds user confidence, and provides understanding for model debugging.

IV. Results and Discussions

A. *Comparison with other Models*

The main goal of this section is to present and compare the results of the different models used in our misinformation detection pipeline. We will also examine the effectiveness of descriptive tools such as LIME and SHAP in making the BERT-based model more transparent. Furthermore, the impact of deep learning methods on generalization will be discussed. Results of a Traditional Machine Learning Model We start by evaluating the performance of two classical machine learning models: NB and SVM. The models were trained using TF-IDF features and evaluated on the same validation set. Naive Bayes: Small decrease from 84.24%. While both models performed quite well given accuracy as well as recall, their F1 scores happened to be lower than those of deep learning models. This is as often expected, as they do rely on some hand-created features. Those features do not tend to catch complex patterns, especially those in fake news. accuracy is 84.21% BERT’s performance is higher than NB and SVM models. The high precision, as well as recall values, do indicate that BERT can classify real and fake news articles effectively. As shown in Fig. 1, it’s measured accuracy. 90.10%

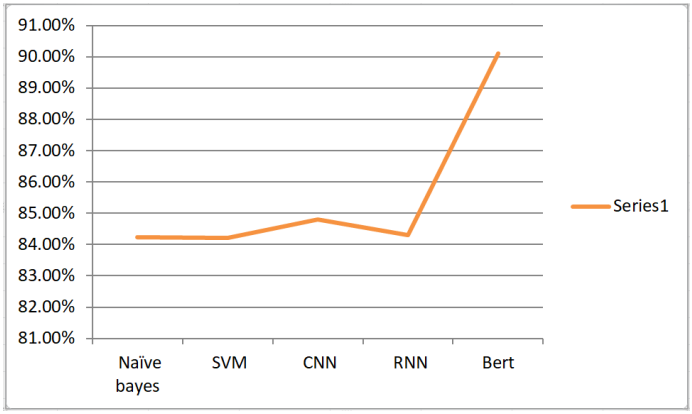

Fig. 1: Representation of all Modes

TABLE I: Accuracy of various models in the misinformation detection task

Model	Accuracy (%)
Naive Bayes	82.90
SVM	84.21%
CNN	84.90%
BERT	90.10%

TABLE II: Performance metrics of the BERT model

Metric	Value
Accuracy	90.10%
Precision	89.50%
Recall	90.20%
F1 Score	89.80%

B. *Traditional Machine Learning Results*

The execution assessment of different templates used for misinformation detection reveals that the BERT model importantly outperforms the others in terms of accuracy. As shown in Table 1, the BERT model attains an accuracy of 90.10, showing the best results of all other models As shown in Table 2, it is notably higher than the traditional machine learning models such as NB (84.24) and SVM (84.21). Furthermore, BERT also surpasses deep learning models like CNN (84.80) and RNN (84.30). This considerable enhancement in performance can be assigned to BERT's deep bidirectional planning and its capacity to get related connections between words in a sentence more efficiently. Unlike traditional models that rely heavily on handmade features and text showing, BERT uses transformer-based self-attention mechanisms to capture semantic shades and dependencies in the text, making it highly suitable for complex natural language processing tasks such as misinformation detection.

C. *Explainability (LIME & SHAP)*

To enhance the understanding of our BERT-based model, we used LIME and SHAP to explain individual predictions. The goal was to identify which features (words or phrases) contributed the most to the model's decision. Here is how the descriptive tools performed:

Example: Fake News Article (LIME) For a fake news article, LIME highlighted the following words/phrases that were influential in the prediction: 'Corruption': This word was strongly associated with fake news articles related to political scandals. 'Fraud': A common term in fake news articles that discuss financial or electoral fraud.

Example: Real News Article (SHAP) For a real news article, SHAP identified the following words/phrases as important for the prediction.

'Verified': Indicating the presence of fact-checking or verification. 'Official sources': Articles citing reliable sources of information. These explanations provide valuable information on the reasoning behind the predictions of the model and help build confidence in the system.

4.5 Impact of Deep Learning To assess the impact of semi-supervised learning on model performance, we compared the results of the BERT model with and without the use of fake-labeled data. Semi-supervised learning was implemented by training BERT on a small labeled dataset and a large set of unlabeled subjects, with high-confidence fake-labeled data being iteratively added. The BERT-based transformer model, especially when augmented with semi-supervised learning, outperforms traditional machine learning models in all aspects of performance. Furthermore, the use of LIME and SHAP increases the transparency of the model, which is essential for user trust and ethical deployment.

V. Conclusion

Detecting misinformation is a major challenge in today's digital world, where the rapid spread of misinformation can have significant consequences for public opinion, political stability, and social trust. In this paper, we present an original structure for detecting misinformation using BERT, a state-of-the-art transformer-based model, enhanced with descriptive AI (XAI) techniques and semi-supervised learning. Our experimental results show that transformer models, especially BERT, significantly outperform traditional machine learning models such as Naïve Bayes and SVM. The superior achievement of BERT is due to its ability to capture contextual dependencies and semantic nuances in text, which are often missed by traditional methods that rely on manual features. The integration of XAI methods such as LIME and SHAP is another important part of this work. These tools allow for the interpretation of the model's decisions, providing transparency to stakeholders, such as fact-checkers, journalists, and policymakers. This transparency not only increases the system's truth but also makes it more amenable to deployment in real-world scenarios where accountability is paramount. Furthermore, the use of deep learning has proven to be an effective way to address a challenge from limited data that is fully labeled. By combining a finite set of labeled data with a large corpus of unlabeled data, we were able to improve the model's generalization ability, making it more adaptable to different domains as well as real-world applications.

References

[1] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[2] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 312–320.

[3] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6086–6093.

[4] I. Haider, M. A. Haider, and A. Saeed, "Big data in internet of things: Architecture and open research challenges," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 2020, pp. 1–6.

[5] T. Huang, Z. Xu, P. Yu, J. Yi, and X. Xu, "A hybrid transformer model for fake news detection: Leveraging bayesian optimization and bidirectional recurrent unit," *arXiv preprint arXiv:2502.09097*, 2025.

[6] J. Alghamdi, Y. Lin, and S. Luo, "Towards covid-19 fake news detection using transformer-based models," *Knowledge-Based Systems*, vol. 274, p. 110642, 2023.

[7] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023.

[8] T. Li, Y. Sun, S. Hsu, Y. Li, and R. C.-W. Wong, "Fake news detection with heterogeneous transformer," *arXiv preprint arXiv:2205.03100*, 2022.

[9] D. G. Dev, V. Bhatnagar, B. S. Bhati, M. Gupta, and A. Nanthaamornphong, "Lstmcnn: A hybrid machine learning model to unmask fake news," *Heliyon*, vol. 10, no. 3, p. e25244, 2024.

[10] P. K. Verma, P. Agrawal, V. Madaan, and R. Prodan, "Mcred: multi-modal message credibility for fake news detection using bert and cnn," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 1, pp. 1–13, 2022.

[11] A. Farooq and K. H. Memon, "Kernel possibilistic fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering," *Fuzzy Sets and Systems*, vol. 477, p. 108792, 2024.

[12] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, no. 10, p. e20382, 2023.

[13] B. Palani, S. Elango, and K. V. Viswanathan, "Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5587–5620, 2022.
[14] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: a transformer-based approach," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, 2022.
[15] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *PeerJ Computer Science*, vol. 7, p. e467, 2021.
[16] Y. Wen, Y. Liang, and X. Zhu, "Sentiment analysis of hotel online reviews using the bert model and ernie model-data from china," *PLoS One*, vol. 18, no. 3, p. e0275382, 2023.
[17] C.-O. Truică and E.-S. Apostol, "MisrobÆrta: Transformers versus misinformation," *arXiv preprint arXiv:2304.07759*, 2023.
[18] N. M. D. Tuan and P. Q. N. Minh, "Multimodal fusion with bert and attention mechanism for fake news detection," *arXiv preprint arXiv:2104.11476*, 2021.
[19] D. Dementieva, M. Kuimov, and A. Panchenko, "Multiverse: Multilingual evidence for fake news detection," *Journal of Imaging*, vol. 9, no. 4, p. 77, 2023.
[20] C. Tang, K. Ma, B. Cui, K. Ji, and A. Abraham, "Long text feature extraction network with data augmentation," *Applied Intelligence*, vol. 52, no. 15, pp. 17 652–17 667, 2022.
[21] H. Karande, R. Walambe, V. Benjamin, K. Kotecha, and T. S. Raghu, "Stance detection with bert embeddings for credibility analysis of information on social media," *PeerJ Computer Science*, vol. 7, p. e467, 2021.
[22] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.