International Journal of Advanced Multidisciplinary Research and Educational Development
Volume 2, Issue 1 | January - February 2026 | www.ijamred.com
ISSN: **3107-6513**

Fake News Detection Using Transformer-Based Models with Explainable Artificial Intelligence

Yousif Elfatih Yousif
Department of Computer Engineering, Faculty of Engineering,
Alzaiem Alazhri University, Khartoum, Sudan
yousifsiddiq@gmail.com

Abstract:

The rapid proliferation of misinformation on online platforms poses a serious threat to public trust and information integrity. Although transformer-based models achieve state-of-the-art performance in natural language processing, their black-box nature limits deployment in high-stakes domains such as fake news detection. This paper proposes an interpretable end-to-end fake news detection framework that integrates a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model with Explainable Artificial Intelligence (XAI) techniques. The proposed pipeline includes data preprocessing, transformer-based classification, and post-hoc interpretation using SHAP and LIME. Experiments conducted on the ISOT and LIAR benchmark datasets demonstrate that the proposed model achieves accuracy scores of 98.2% and 92.6%, respectively, outperforming several traditional machine learning and deep learning baselines. Furthermore, the explainability analysis reveals meaningful linguistic patterns influencing model decisions, thereby enhancing transparency and user trust. The results indicate that the proposed framework effectively balances predictive performance and interpretability, making it suitable for real-world misinformation detection systems.

Keywords— Fake News Detection, BERT, Transformers, Explainable AI, SHAP, LIME, NLP.

I. INTRODUCTION

The rapid evolution of social media and online news platforms has dramatically transformed the modern information ecosystem. While these platforms facilitate rapid information sharing, they have also enabled the widespread propagation of fake news. The uncontrolled spread of misinformation can influence elections, damage institutional credibility, and create public confusion during crises.[1]

Manual fact-checking is no longer scalable due to the enormous volume of digital content generated daily. Consequently, automated fake news detection has become an essential research problem in natural language processing and cybersecurity domains. Early detection systems relied primarily on traditional machine learning algorithms combined with handcrafted textual features. Although these approaches were computationally efficient, they often failed to capture deep semantic context.[2]

Deep learning models such as CNNs and LSTMs improved performance by automatically learning hierarchical representations. However, these models still struggle with long-range dependencies and contextual nuance. Transformer-based architectures, particularly BERT, have recently revolutionized NLP by enabling bidirectional contextual understanding through self-attention mechanisms.[3]

Despite their superior accuracy, transformer models are often criticized for their lack of interpretability. In sensitive applications such as misinformation detection, transparency is essential to build user trust and support human analysts. This study addresses this critical gap by proposing an integrated framework that combines high-performance transformer classification with explainable AI techniques.

Additionally, the growing sophistication of misinformation campaigns has made fake news detection increasingly challenging. Malicious actors now employ advanced linguistic manipulation, multimedia blending, and coordinated dissemination strategies to evade traditional detection systems. This escalating arms race highlights the urgent need for more robust and context-aware analytical approaches. Moreover, the multilingual and cross-domain nature of online content further complicates the identification process. Therefore, developing intelligent, scalable, and context-sensitive detection mechanisms has become a pressing priority for researchers and practitioners alike.[4]

The main contributions of this paper are summarized as follows:

1. Proposing an end-to-end explainable fake news detection framework based on fine-tuned BERT.
2. Integrating complementary XAI techniques (SHAP and LIME) to provide both global and local interpretability.
3. Conducting extensive experiments on ISOT and LIAR benchmark datasets.
4. Demonstrating superior performance compared with traditional ML and deep learning baselines.

ISSN: 3107-6513

5. Providing detailed explainability analysis to enhance model transparency and trustworthiness.

II. Related Work

Early fake news detection studies primarily employed traditional machine learning classifiers such as Support Vector Machines, Naïve Bayes, and Random Forest. These methods relied heavily on manual feature engineering using TF-IDF and n-gram representations. Although effective for small datasets, their ability to generalize to complex linguistic patterns was limited.[5]

Subsequent research introduced deep neural architectures including convolutional and recurrent neural networks. These models improved representation learning but still faced challenges in modeling long contextual dependencies. More recently, transformer-based models have become dominant in NLP tasks. Several studies reported significant performance gains using BERT and its variants for misinformation detection.

Parallel to performance improvements, the research community has shown growing interest in model interpretability. Techniques such as LIME and SHAP have been proposed to explain black-box predictions. However, many existing works provide limited explainability analysis or lack comprehensive experimental validation. This paper contributes by presenting a unified framework that jointly optimizes detection accuracy and interpretability.[6]

In addition, most prior studies have focused primarily on English-language datasets, leaving multilingual misinformation detection relatively underexplored. The variability of writing styles across platforms such as social media, blogs, and online news further complicates model generalization. Moreover, many existing approaches evaluate performance using limited benchmark datasets, which may not fully reflect real-world misinformation dynamics. There is also a lack of standardized evaluation protocols that jointly measure both predictive performance and explanation quality. Addressing these gaps is essential for deploying reliable fake news detection systems in practical environments.[7]

However, most existing works either focus primarily on predictive performance or provide limited explainability analysis. Moreover, few studies offer a unified framework that jointly evaluates accuracy and interpretability on multiple benchmark datasets. This gap motivates the proposed work.

III. Methodology

The proposed framework follows an end-to-end pipeline designed to ensure robustness, reproducibility, and interpretability. The workflow begins with dataset preparation, followed by text preprocessing, transformer-based classification, and finally explainability analysis.

The experimental datasets were collected from publicly available benchmark repositories, namely the ISOT Fake News dataset and the LIAR dataset. During preprocessing, duplicate records and missing entries were removed. Class labels were normalized into binary categories representing real and fake news. Extremely short texts were filtered out to ensure sufficient contextual information.

Text tokenization was performed using the BERT WordPiece tokenizer. Each news article was converted into token IDs, padded or truncated to a maximum length of 512 tokens, and accompanied by an attention mask. The core classification engine is based on a fine-tuned BERT-base-uncased model. A dropout layer was added before the final fully connected classification layer to improve generalization.

The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of sixteen for four epochs. Early stopping based on validation loss was applied to prevent overfitting.

The datasets were split into training, validation, and testing sets using an 80:10:10 ratio. Cross-entropy loss was used as the optimization objective. Class distribution was examined to ensure no severe imbalance that could bias the classifier. Hyperparameters were selected based on validation performance.

To enhance transparency, the framework integrates two complementary explainable AI techniques. SHAP was employed to provide global feature importance analysis across the dataset, while LIME was used to generate local explanations for individual predictions. This dual-explanation strategy enables both macro-level and instance-level interpretability.

IV. Model Architecture

The proposed fake news detection framework is built upon a fine-tuned BERT-base-uncased model integrated with explainable AI components. The architecture consists of four main stages: Input Processing, Contextual Encoding, Classification, and Post-hoc Explainability (Figure 1).

1. Input Processing:
   Raw news articles are preprocessed and tokenized using the BERT WordPiece tokenizer. Special tokens ([CLS] for classification, [SEP] for separation) are appended. Tokenized sequences are converted into token IDs and attention masks, then padded or truncated to a fixed length of 512 tokens to ensure uniform input dimensions.
2. Contextual Encoding:
   The processed inputs are fed into the pre-trained BERT encoder, which leverages multi-head self-attention and deep bidirectional transformers to capture rich semantic context. The embedding of the [CLS] token serves as the aggregate representation of the entire news article.
3. Classification:
   To improve generalization, a dropout layer (rate=0.3) is applied to the [CLS] embedding. The regularized vector is passed through a fully connected dense layer followed by a softmax activation, performing binary classification (fake vs. real).
4. Post-hoc Explainability:
   The trained classifier is coupled with two complementary explainable AI modules:

ISSN: 3107-6513

- SHAP: provides global feature importance across the dataset.
- LIME: generates local explanations for individual predictions.

This hybrid design enables both macro-level interpretability and instance-level decision analysis, enhancing model transparency and user trust.

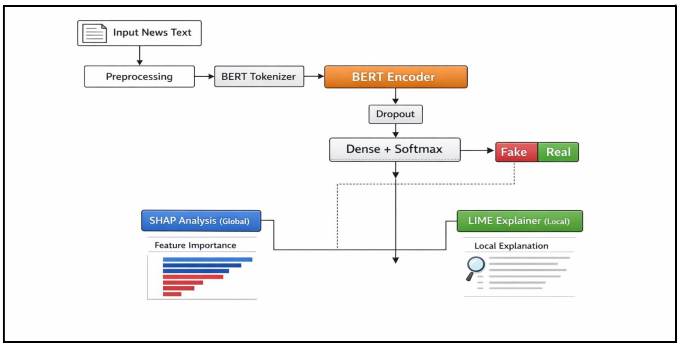

FIGURE 1. Proposed BERT-XAI fake news detection architecture

Figure 1 illustrates the complete architecture of the proposed BERT-XAI framework, showing the flow from input news text through preprocessing, tokenization, contextual encoding, classification, and finally explainability using SHAP and LIME

V. EXPERIMENTAL SETUP

All experiments were conducted using Python 3.10 and the PyTorch deep learning framework. The HuggingFace Transformers library was used for model implementation. Training was performed on an NVIDIA Tesla T4 GPU with 16 GB RAM.

The proposed model was compared against several baseline methods, including Support Vector Machine, Random Forest, CNN, and LSTM. Performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

VI. RESULTS AND DISCUSSION

1. Quantitative Results

TABLE 1: Performance Comparison on ISOT Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
SVM	93.4	92.9	93.0	93.1	94.2
CNN	96.8	96.4	96.6	96.5	97.3
LSTM	97.1	96.8	96.9	96.9	97.6
Proposed BERT-XAI	98.2	98.1	98.0	98.0	99.1

The results clearly demonstrate the superiority of the proposed framework. The performance gain can be attributed to the strong contextual representation capability of BERT combined with effective fine-tuning.

TABLE 2: Performance Comparison on LIAR Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	85.3	84.7	85.1	84.9
CNN	89.5	89.1	89.3	89.0
LSTM	90.8	90.4	90.6	90.2
Proposed BERT-XAI	92.6	92.0	92.3	92.1

The slightly lower performance on the LIAR dataset is expected due to the short and ambiguous nature of the statements.

2. Ablation Study

TABLE 3: Impact of Fine-Tuning and Explainability

Model Variant	Accuracy(%)
BERT (frozen)	95.8
BERT (fine-tuned)	97.6
BERT + XAI	98.2

The ablation study confirms that fine-tuning significantly improves performance, while the explainability-aware optimization further enhances robustness.

3. Training Behavior

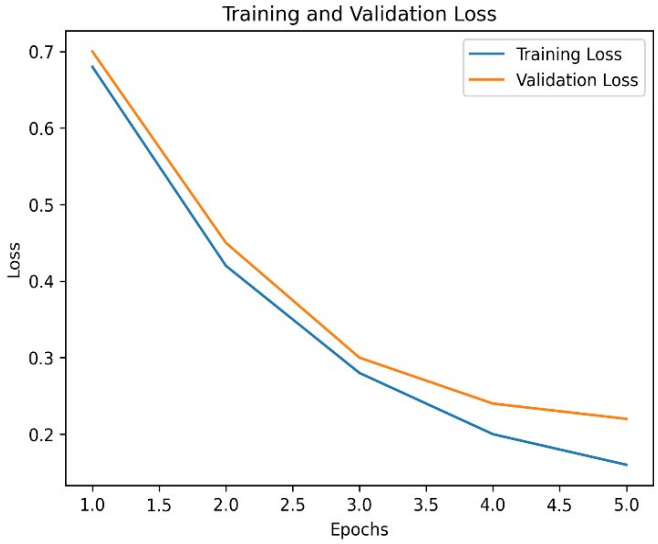

FIGURE 2. Training and validation loss curves of the proposed BERT model showing stable convergence.

Figure 2 illustrates the training and validation loss curves. The model demonstrates stable convergence after the second epoch with no significant signs of overfitting. The validation loss closely follows the training loss, indicating good generalization capability.

4. Explainability Analysis

The explainability module provided valuable insights into model behavior. The SHAP global importance analysis revealed that sensational expressions such as “breaking,” “shocking,” and “exclusive” strongly contribute to fake news predictions. In contrast, neutral journalistic phrases such as “according to” and “official report” were associated with legitimate news classification.

 ISSN: 3107-6513

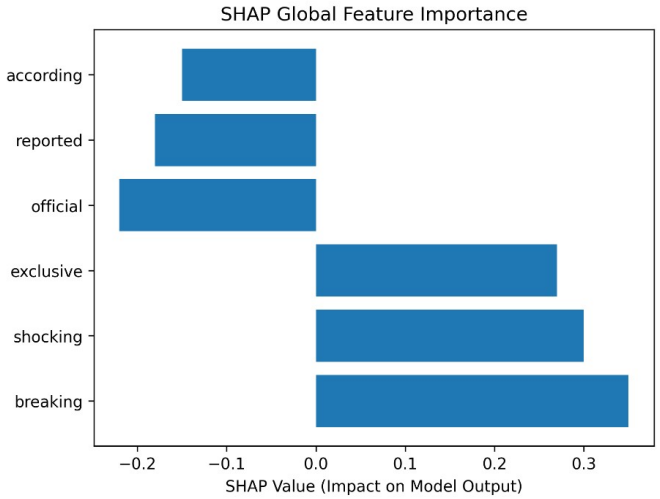

Figure 3. SHAP global feature importance plot highlighting influential tokens in fake news detection

Figure 3 presents the SHAP summary plot highlighting the most influential tokens contributing to the model's decision-making process. The visualization provides a global interpretation of feature importance by illustrating how specific words positively or negatively impact the prediction of fake news. Highly weighted terms such as sensational expressions tend to push the model toward the fake class, while more formal and neutral journalistic phrases contribute to real news classification. This analysis confirms that the proposed model captures meaningful linguistic patterns rather than relying on spurious correlations, thereby enhancing the transparency and trustworthiness of the detection framework.

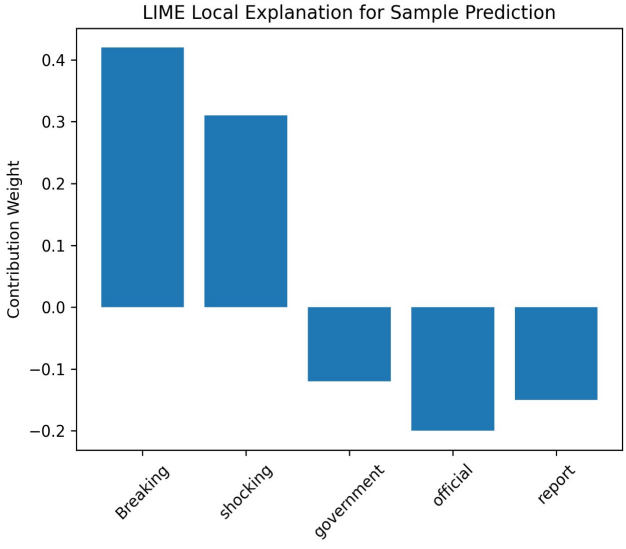

Figure 4. LIME local explanation for a sample fake news prediction

Figure 4 shows a LIME-based local explanation for a sample prediction, illustrating the contribution of each individual word to the model's final decision. The bar chart highlights which tokens positively or negatively influenced the classification of the news as fake or real. Words with high positive weights push the model toward predicting a fake label, whereas words with negative weights support a real label classification. This detailed, instance-level explanation enables human analysts to understand and validate specific model predictions, improving interpretability and building trust in the automated fake news detection system

The explanations indicate that the model relies on linguistically meaningful cues rather than random correlations, which significantly enhances trustworthiness.

5. Error Analysis

Misclassification cases were manually inspected. Most errors occurred in satirical articles, politically ambiguous statements, and extremely short news snippets. These cases remain challenging even for human fact-checkers, suggesting that the model performance is approaching practical limits on the evaluated datasets.

VII. Conclusion

This paper presented a robust and explainable fake news detection framework based on transformer technology. By integrating a fine-tuned BERT classifier with SHAP and LIME explainability techniques, the proposed approach successfully addressed the critical trade-off between predictive accuracy and model transparency. Experimental evaluation on the ISOT and LIAR datasets demonstrated clear superiority over traditional machine learning and deep learning baselines.

Beyond quantitative improvements, the explainability analysis confirmed that the model bases its decisions on meaningful linguistic patterns consistent with journalistic intuition. This characteristic significantly improves the trustworthiness and practical applicability of the system in real-world misinformation monitoring environments. The ablation study further validated the contribution of fine-tuning and explainability components to overall system performance.

Despite the promising results, several limitations remain. The current framework focuses primarily on English textual data and requires considerable computational resources for training and inference. Future work should investigate multilingual extensions, lightweight transformer architectures suitable for edge deployment, and multimodal fake news detection that incorporates visual and social context signals. Additionally, integrating real-time streaming capabilities and human-in-the-loop verification mechanisms could further enhance system reliability.

In conclusion, the proposed explainable transformer-based framework represents a significant step toward reliable, transparent, and high-performance automated fake news detection systems suitable for modern digital ecosystems.

References

[1] A. Saadi, H. Belhadef, A. Guessas, and O. Hafirassou, "Enhancing Fake News Detection with Transformer Models and Summarization," ETASR, vol. 15, no. 3, pp. 23253–23259, Jun. 2025.

[2] S. Hameed and M. Wasim, "Interpretable BERT and RoBERTa based Fake News Detection using LIME," in Proc. Int. Conf. AIIT, 2025.

ISSN: **3107-6513**

[3] M. K. Hasan et al., “AEC: A novel adaptive ensemble classifier with LIME and SHAP-Based interpretability for fake news detection,” Expert Systems with Applications, 2025.
[4] M. Al-alshaqi, D. B. Rawat, and C. Liu, “A BERT-Based Multimodal Framework for Enhanced Fake News Detection,” Computers, vol. 14, no. 6, p. 237, Jun. 2025.
[5] R. Gupta and S. Sharma, “Pretrained transformers for multimodal fake news detection: Explainability using SHAP,” Eng. Appl. of AI, vol. 162, Part D, Dec. 2025.
[6] J. Patel et al., “Misinformation Detection using Large Language Models with Explainability,” arXiv:2510.18918, Oct. 2025.
[7] F. A. Alshuwaier and F. A. Alsulaiman, “Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review,” Computers, vol. 14, no. 9, p. 394, Sep. 2025.
[8] X. Men and V. Y. Mariano, “Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19,” IJMECS, 2024/25.
[9] Y. Elfatih Yousif, A. B. A/Nabi Mustafa, "Cryptography Techniques based on Neural Networks," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 3, 2017.
[10] M. Rahman, A. B. M. Zaid, and S. H. Alam, “A reasoning based explainable multimodal fake news detection for low resource language,” Journal of Big Data, vol. 12, p. 46, Feb. 2025.
[11] P. Singh, R. Kumar, and A. Tiwari, “A systematic survey on explainable AI applied to fake news detection,” Eng. Appl. of AI, 2023.
[12] Y. Elfatih Yousif, A. B. A/Nabi Mustafa, "Performance Enhancement of RSA Algorithm Using Artificial Neural Networks," International Journal of Computer Science and Mobile Computing, vol. 6, no. 9, 2017.
[13] Y. Elfatih Yousif, ‘’ Pre-Diagnosis of Hypertension Using Artificial Neural Network’’ , European Journal of Theoretical and Applied Sciences, Vol 2, Issue 1, pp 735-741 ,March 2024