An Integrated LLM and XAI Approach to Misinformation Detection

Wan Nursyafeeza Binti Wan Mohd Nasir, Chimeremma Sandra Amadi, Taesoo Jun
(Department of Software Engineering, Department of IT Convergence Engineering)
Kumoh National Institute of Technology Gumi, South Korea
wsyafeeza@yahoo.com, chimesandra@yahoo.com, taesoo.jun@kumoh.ac.kr

***Abstract*—Misinformation poses significant risks to democracy and public safety. An misinformation detection framework that combines Explainable AI (XAI), traditional classifiers, and Large Language Models (LLMs) like BERT and GPT is presented in this paper. News feeds are processed, characteristics are extracted and texts are categorized with justifications for each classification. Analysis reveals that BERT outperforms TF-IDF (Term Frequency-Inverse Document Frequency) with Logistic Regression in terms of F1-score, Accuracy, Precision, and Recall. XAI tools like SHAP and LIME improve trust and transparency. This method enhances interpretability and detection performance.**

***Index Terms*—Large Language Models (LLMs), Natural Language Processing (NLP), Misinformation Detection, Explainable AI (XAI)**

I. Introduction

Misinformation, which is defined as false or misleading information presented as fact, has become a major problem due to the growing use of social media and online news sources [1]. It affects political outcomes, undermines public confidence, and poses a threat to public health and safety [2]. As a result, automatic misinformation detection has emerged as a important research topic.

LLMs such as Google's BERT and OpenAI's GPT [3] have demonstrated exceptional performance in a variety of Natural Language Processing (NLP) tasks, such as text generation, question answering, and language translation [4], [5]. LLMs are frequently utilized in applications such as chatbots, translation tools, and text summarization because they are trained to understand and generate human language.

Despite their achievements, LLMs also present challenges, particularly the potential to create convincing but false content that unintentionally spreads misinformation [6], [7]. In this paper, a framework for misinformation detection system that combines LLM-based feature extraction, XAI techniques, and traditional classifiers is proposed. The system not only classifies news articles as "True" or "Misinformation", but also provides explanations to support transparency and accountability.

The following are the key contributions of this study:

- A proposed of explanation generation in misinformation detection to improve user trust and system reliability.
- Providing a detailed analysis of how LLMs can enhance misinformation detection by identifying and minimizing the spread of false information.

II. Proposed Methodology

The proposed methodology for the misinformation detection system is described in this section. News feed ingestion, web scraping, data collection, dataset compilation, data processing, feature extraction, LLM-based classification, misinformation detection, explanation generation, and output delivery are some of the main steps of the methodology as illustrated in Fig. 1. These will be covered in more detail in the following sections.

The process begins with collecting data from various news sources via web scraping, RSS feeds, and APIs. Every entry has metadata including the full content, publication date, source and headline. After unnecessary entries are eliminated, the data is organized into a raw dataset. Simple normalization is used including deleting HTML and lowercasing text. Lastly, articles are classified as either "True" or "Misinformation" in order to provide ground truth for model training.

Several pre-processing steps are performed on the raw textual data before machine learning models are applied. This includes text cleaning to remove duplicate content, punctuation, special characters and non-informative elements, tokenization to split text into individual tokens or words, stopword removal to eliminate common words that do not contribute to the analysis, stemming to reduce words to their root forms to unify variations and synonym normalization to map synonymous terms to a single canonical form to ensure feature consistency. To convert textual input into organized numerical characteristics appropriate for machine learning, feature extraction is carried out after the data has been cleaned and normalized. Word embeddings such as GloVe may be used to capture semantic meaning, while TF-IDF is used to evaluate word importance across documents. To assess model performance, the dataset is then divided into training and testing sets. This is to make sure the algorithm can effectively generalize to unseen news articles.

An LLM-based classification model is used at this fundamental stage to evaluate the features that have been retrieved and assign a "True" or "Misinformation" label to each article. The labeled datasets can be used to refine pre-trained LLMs, such as GPT and BERT. Two models are compared, which are TF-IDF with LR and BERT Transformer. Both models were tested on labeled datasets, GossipCop and Politifact [8]–[10]. The performance of each model was evaluated based on

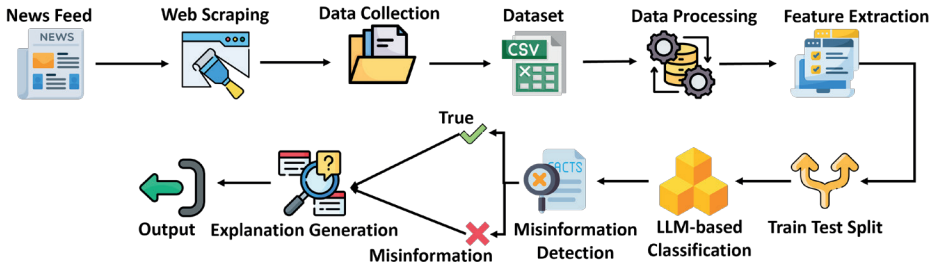

Fig. 1: Illustration of misinformation detection workflow.

Accuracy, Precision, Recall and F1-score. The results are as shown in Table I.

TABLE I: Comparison of TF-IDF + Logistic Regression with BERT Transformer.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
TF-IDF + LR	84.98	85.48	96.55	90.68
BERT Transformer	86.34	89.84	92.31	91.06

The efficiency is maintained over several evaluation runs, even though BERT only slightly outperforms the traditional model in terms of overall accuracy and class balance. Despite the slight numerical discrepancy, BERT shows better robustness and generalization. Furthermore, BERT continues to show potential in situations where model correctness and fairness are crucial, even after considering for inference time and computational cost.

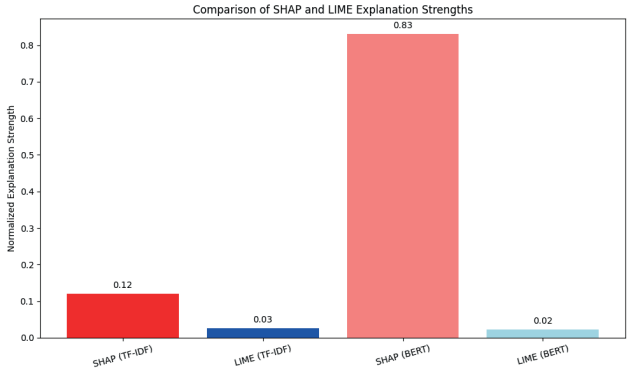

Fig. 2: Comparison of SHAP&LIME on TF-IDF&BERT

XAI techniques like SHAP and LIME are utilized to increase transparency and user trust by identifying which features influence each predictions, helping users understand why an article is flagged as ”True” or ”Misinformation”. According to the graph in Fig. 2, LIME performs significantly worse across both BERT and TF-IDF representations, but SHAP with BERT embeddings provides the best explanations which is 0.83. Overall, SHAP offers more insightful and useful explanations especially when combined with BERT.

III. Conclusion

The proposed framework combines XAI techniques, traditional models and LLMs to detect misinformation. According to experiments, BERT outperforms TF-IDF with LR in both accuracy and metric balance. The system builds user trust and transparency by explaining each result by using SHAP or LIME. It helps users by determining the credibility of the news articles. Future work may improve explanation quality and apply more advanced techniques.

Acknowledgment

This research was funded by the Innovative Human Resource Development for Local Intellectualization Program (IITP-2025-RS-2020-II201612, 33%) through IITP under MSIT, the Basic Science Research Program (2018R1A6A1A03024003, 33%) through NRF, and the Information Technology Research Center (ITRC) Program (IITP-2025-RS-2024-00438430, 34%) funded by MSIT through IITP.

References

[1] C. Chen and K. Shu, “Can llm-generated misinformation be detected?” 2024.

[2] T. Huang, J. Yi, P. Yu, and X. Xu, “Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies,” 2025.

[3] “Summary of chatgpt-related research and perspective towards the future of large language models,” *Meta-Radiology*, vol. 1, no. 2, p. 100017, 2023.

[4] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, “A review on large language models: Architectures, applications, taxonomies, open issues and challenges,” *IEEE Access*, 2024.

[5] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 17 061–17 084.

[6] C. Chen and K. Shu, “Combating misinformation in the age of llms: Opportunities and challenges,” *AI Magazine*, vol. 45, no. 3, pp. 354–368, 2024.

[7] P. Santra, “Leveraging llms for detecting and modeling the propagation of misinformation in social networks,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 3073.

[8] K. Shu, S. Wang, and H. Liu, “Exploiting tri-relationship for fake news detection,” *arXiv preprint arXiv:1712.07709*, 2017.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[10] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media,” *arXiv preprint arXiv:1809.01286*, 2018.