Generative AI-Driven Fake News Detection and Rewriting Framework Using GPT for Accurate Media Verification and Content Correction

[1]B. Purna Chandra Babu, [2]N. Madhava Rao
[1]M. Tech Student, Department of CSE, Prakasam Engineering College (Autonomous), Kandukur, India
[2]Associate Professor, Department of CSE, Prakasam Engineering College (Autonomous), Kandukur, India

ABSTRACT: ***Fake news has become a major challenge in the digital era, spreading misinformation rapidly through social media platforms, news websites, and online forums. Manual fact-checking methods are slow, inconsistent, and unable to cope with the massive volume of online content generated every day. To address this issue, this project proposes a Generative AI-Driven Fake News Detection and Rewriting Framework using GPT-based models. The system automatically analyzes news articles, social media posts, and textual content to detect misinformation using Natural Language Processing (NLP) and machine learning techniques. Once fake or misleading content is identified, the framework generates a corrected and fact-aligned version of the news using generative AI. This approach not only identifies fake news but also assists in content correction and media verification. The proposed framework aims to improve the reliability of online information, reduce human effort, and support responsible information dissemination in digital media.***

KEYWORDS: ***- Fake News Detection, Generative AI, GPT, Natural Language Processing, Machine Learning, Transformer Models, Media Verification, Text Rewriting.***

INTRODUCTION

Fake news has become a serious issue in the digital communication era. The rapid growth of the internet and social media platforms allows information to spread quickly among users. However, not all online information is genuine or verified. Misleading news can influence public opinion, create unnecessary panic, damage reputations, and cause social problems.

Traditional fact-checking methods mainly depend on manual verification. These methods require significant time and human effort, making it difficult to analyze the large amount of content shared every day. Basic detection systems can classify news

as real or fake, but they generally do not provide corrected information to users.

Artificial Intelligence and Natural Language Processing provide effective techniques for analyzing textual content. Transformer-based models can understand contextual relationships in sentences, while GPT-based generative models can produce meaningful and context-aware text.

LITERATURE REVIEW

Several research approaches have been developed to identify fake news using machine learning, deep learning, and Natural Language Processing techniques.

Traditional machine learning methods analyze textual features and classify news content based on previously learned patterns. These approaches are useful for identifying known misinformation patterns, but they may have limitations when the writing style or context changes.

Deep learning approaches improve text analysis by automatically learning important features from textual data. Transformer-based models provide better contextual understanding by analyzing relationships between words and sentences. Generative AI models introduce additional capabilities by generating human-readable text. In the proposed project, generative AI is used not only for identifying misleading information but also for producing corrected and fact-aligned content.

However, many existing approaches mainly focus on classification. The proposed framework combines fake news detection, factual verification, explanation, and content rewriting in one system.

RELATED WORK

The proposed project uses Python for the core processing and integrates Natural Language Processing and transformer-based techniques for fake news detection. The system receives textual content from the user interface and performs preprocessing operations such as tokenization, cleaning, and normalization.

The processed text is analyzed using a transformer-based classification approach. BERT or RoBERTa are discussed as possible models in the source project. The classification module identifies the submitted content as real or fake and provides a probability score.

When misleading content is detected, the GPT-based rewriting component generates corrected text while preserving the original topic and context. Semantic similarity and factual verification techniques help maintain content consistency. The framework also provides explanations and visualizations, enabling users to understand detection results, identify misleading information, and interpret the generated corrections more effectively.

EXISTING METHOD

Existing fake news detection systems mainly depend on manual fact-checking, keyword-based filtering, and traditional machine learning classification techniques. Manual verification requires human reviewers to examine news articles and compare information with reliable sources. This process is time-consuming and difficult to manage when large volumes of content are generated.

Some automated systems use predefined keywords and textual patterns to identify suspicious information. However, these methods may fail to understand the actual context of a sentence. Traditional machine learning and deep learning systems can classify content as real or fake. However, many of these systems only provide a classification result and do not generate corrected information.

These limitations create a need for an integrated framework that combines detection, verification, and rewriting.

PROPOSED METHOD

The proposed system introduces a Generative AI-Driven Fake News Detection and Rewriting Framework using GPT-based models. The system accepts news articles, social media posts, or other textual content from the user. Natural Language Processing techniques are applied to clean and prepare the input text. The processed content is then analyzed using a transformer-based classification approach.

The detection module identifies whether the submitted content is real or fake and provides a probability score. If the content is identified as misleading, the GPT-based rewriting module generates a corrected version of the text. The framework also considers factual verification using reliable information sources and semantic consistency checks. Explainability features help users understand why the content has been identified as suspicious.

SYSTEM ARCHITECTURE

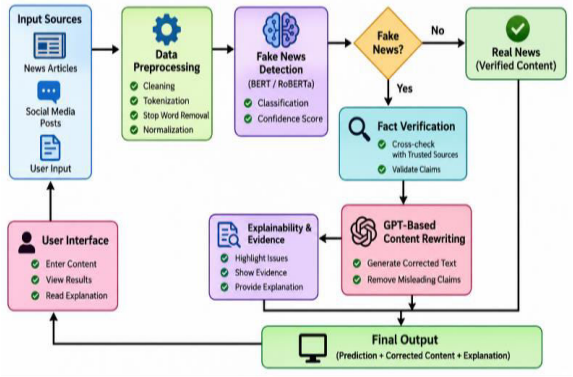

Fig. 1. System architecture of the proposed fake news detection and rewriting framework.

The proposed system follows a sequential processing architecture that integrates data collection, preprocessing, fake news detection, fact verification, GPT-based rewriting, and explainability. The system first receives textual content from the user or supported data sources. The preprocessing module cleans the input and prepares it for analysis. The detection

engine uses transformer-based models such as BERT or RoBERTa to identify misleading content and generate probability scores.

The Fact Verification Module compares suspicious claims with trusted sources and knowledge bases. If misinformation is identified, the GPT-Based Content Rewriting Module generates a corrected version of the content. Finally, the Explainability and Evidence Generation Layer highlights misleading portions and displays supporting evidence.

METHODOLOGY DESCRIPTION

1. Data Collection Module: This module collects textual data from news websites, social media platforms, blogs, and user submissions through APIs and labeled datasets.

2. Data Preprocessing Module: This module cleans the collected text by removing unwanted symbols, tokenizing words, removing stop words, and applying stemming or lemmatization.

3. Feature Extraction Module: This module converts text into numerical representations using word embeddings and contextual embeddings to identify meaningful language patterns.

4. Fake News Detection Module: This module uses transformer-based models such as BERT or RoBERTa to classify news content as real or fake and generate confidence scores.

5. Fact Verification Module: This module cross-checks suspicious claims with trusted sources and knowledge bases to evaluate the factual accuracy of the content.

6. GPT-Based Content Rewriting Module: This module rewrites misleading information into clear and corrected text while maintaining the original topic and context.

7. Explainability Module: This module highlights misleading text and displays prediction confidence, supporting evidence, and explanations to help users understand the results.

8. User Interface Module: This module allows users to enter or upload news content and view detection results, confidence scores, explanations, and rewritten content.

RESULTS AND DISCUSSION

Figures 2 to 4 show the implemented web application of the proposed framework, covering the home page, the news input page and a sample detection result.

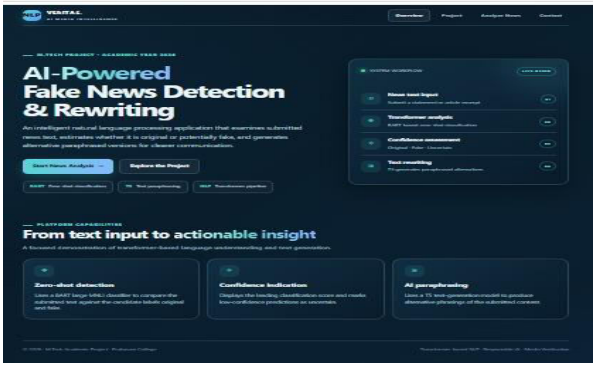

Fig. 2. Home page of the Veritas application.

Landing page that introduces zero-shot detection, confidence indication and AI paraphrasing as the three main capabilities.

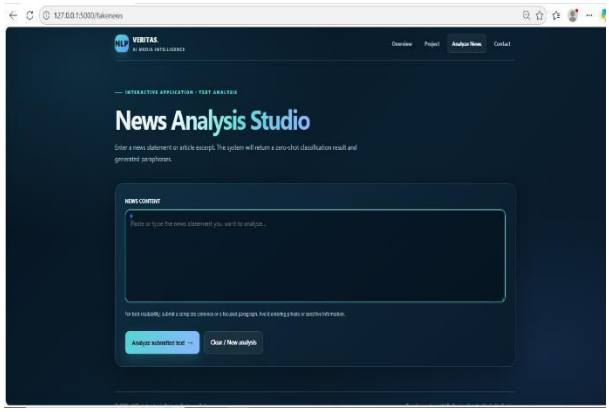

Fig. 3. News Analysis Studio (input page).

Text box where the user pastes a news statement or excerpt, with Analyze and Clear buttons.

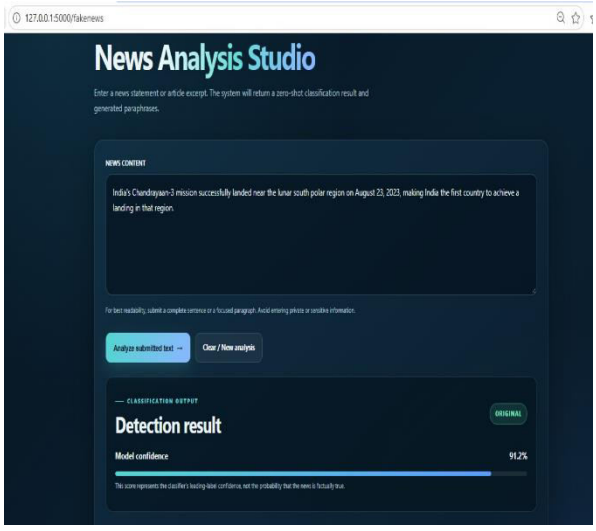

Fig. 4. Detection result for a sample news statement.

The Chandrayaan-3 landing statement is classified as ORIGINAL with a model confidence of 91.2%.

The confidence value shown in Fig. 4 is the classifier's leading-label confidence and not the probability that the news is factually true. Quantitative evaluation (accuracy, precision, recall and F1-score) and sample outputs of the rewriting module are not included in the supplied material and should be added before final submission.

CONCLUSION

The proposed Generative AI-Driven Fake News Detection and Rewriting Framework combines NLP, transformer-based classification, fact verification, and GPT-based content generation. It identifies fake news, verifies suspicious claims using trusted sources, and rewrites misleading content while preserving its original context. Explainability features provide transparency through prediction results and supporting evidence. Overall, the framework integrates fake news detection, verification, and content correction into a unified system.

FUTURE SCOPE

The proposed system can be enhanced by introducing multilingual fake news detection to support different languages. Future improvements may include real-time monitoring of social media platforms, integration with advanced knowledge bases for better fact verification, and image and video misinformation detection. The system can also be improved through continuous model updates, enhanced explainability, and integration with news platforms to provide faster and more reliable misinformation detection and content correction.

REFERENCES

1. 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, pp. 335–362, 2022. https://doi.org/10.1007/s41060-021-00302-z

2. F. Folino, G. Folino, M. Guarascio, L. Pontieri, and P. Zicari, "Towards data- and compute-efficient fake-news detection: An approach combining active learning and pre-trained language models," *SN Computer Science*, vol. 5, Art. no. 470, 2024. https://doi.org/10.1007/s42979-024-02809-1

3. S. Kuntur, M. Krzywda, A. Wróblewska, M. Paprzycki, and M. Ganzha, "Comparative analysis of graph neural networks and transformers for robust fake news detection: A verification and reimplementation study," *Electronics*, vol. 13, no. 23, Art. no. 4784, 2024. https://doi.org/10.3390/electronics13234784

4. E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A survey on the use of large language models (LLMs) in fake news," *Future Internet*, vol. 16, no. 8, Art. no. 298, 2024. https://doi.org/10.3390/fi16080298

5. J. Su, C. Cardie, and P. Nakov, "Adapting fake news detection to the era of large language models," in *Findings of the Association for Computational Linguistics: NAACL*, 2024, pp. 1473–1490.

6. J. Wang, Z. Zhu, C. Liu, R. Li, and X. Wu, "LLM-enhanced multimodal detection of fake news," *PLOS ONE*, vol. 19, no. 10, Art. no. e0312240, 2024. https://doi.org/10.1371/journal.pone.0312240

7. M. Al-alshaqi, D. B. Rawat, and C. Liu, "Ensemble techniques for robust fake news detection: Integrating transformers, natural language processing, and machine learning," *Sensors*, vol. 24, no. 18, Art. no. 6062, 2024. https://doi.org/10.3390/s24186062

8. S. Harris, H. J. Hadi, N. Ahmad, and M. A. Alshara, "Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas," *Technologies*, vol. 12, no. 11, Art. no. 222, 2024. https://doi.org/10.3390/technologies12110222

9. K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models," *Future Internet*, vol. 17, no. 1, Art. no. 28, 2025. https://doi.org/10.3390/fi17010028

10. E. Papageorgiou, I. Varlamis, and C. Chronis, “Harnessing large language models and deep neural networks for fake news detection,” *Information*, vol. 16, no. 4, Art. no. 297, 2025. https://doi.org/10.3390/info16040297

11. J. Rout, M. Mishra, and M. J. Saikia, “Towards reliable fake news detection: Enhanced attention-based transformer model,” *Journal of Cybersecurity and Privacy*, vol. 5, no. 3, Art. no. 43, 2025. https://doi.org/10.3390/jcp5030043

12. H. R. Iyer and A. K. Madasamy, “A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers,” *Journal of Big Data*, vol. 12, Art. no. 46, 2025. https://doi.org/10.1186/s40537-025-01093-x