IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

"An Explainable Al Framework for Multimodal Misinformation Detection and Credibility Assessment"

Ms. Ashwini Baldev Khairnar[1], Prof. Shivam Limbhare[2], Prof. Manoj Vasant Nikum[3]

Research Scholar, Master of Computer Applications, Shri JRIT, Dondaicha, KBCNMU Jalgaon, Maharashtra, India[1]

Assistant Professor, Master of Computer Applications, Shri JRIT, Dondaicha, KBCNMU Jalgaon, Maharashtra, India[2]

HOD, Master of Computer Applications, Shri JRIT, Dondaicha, KBCNMU Jalgaon, Maharashtra, India[3]

Abstract: The rapid growth of digital media and social networking platforms has significantly increased the spread of misinformation, creating challenges for individuals, organizations, and society. Traditional misinformation detection systems often rely on limited textual features and provide binary predictions without explaining the reasons behind their decisions. This project proposes an Explainable Artificial Intelligence (XAl) framework for multimodal misinformation detection and credibility assessment. The proposed framework is designed to analyze multiple forms of information, including textual content and, where applicable, associated visual or contextual features, to identify patterns that may indicate misleading or unreliable information. Natural Language Processing (NLP), machine learning, transformer-based semantic representations, and multimodal feature extraction are combined to improve detection performance. In addition to classifying content as credible or potentially misleading, the framework generates a credibility score and provides interpretable explanations showing the major factors that influenced the prediction. Explainability techniques are incorporated to reduce the black-box nature of Al-based decision-making and improve user trust. The proposed system can be implemented as an interactive application that accepts user-submitted content and presents the predicted classification, confidence level, credibility assessment, and supporting explanations. The framework can be evaluated using standard performance measures such as accuracy, precision, recall, F1-score, and explainability-related analysis. The overall objective is to develop a transparent, scalable, and research-oriented Al solution that assists users in assessing potentially misleading information while providing understandable evidence for the system's decisions.

Keywords: Explainable Al, Misinformation Detection, Multimodal Al, Natural Language Processing, Credibility Assessment, Transformer Models, Machine Learning, Deep Learning.

I. INTRODUCTION

The rapid advancement of digital technologies, internet connectivity, and social networking platforms has transformed the way information is created, shared, and consumed. Platforms such as social media, online news portals, blogs, and digital communication networks allow users to access and distribute information within seconds. Although these technologies provide significant benefits in communication and knowledge sharing, they have also created serious challenges related to the spread of misinformation and misleading content.

Misinformation refers to false, inaccurate, or misleading information that is shared regardless of whether the individual spreading it intends to deceive others. The widespread circulation of such information can influence public opinion, create social confusion, damage reputations, and negatively affect decision-making processes. During major events such as elections, public health emergencies, and international conflicts, misinformation can spread rapidly and become difficult to control.

II. LITERATURE SURVEY

2.1 Introduction

The rapid growth of social media, digital communication, and online news platforms has increased the availability and circulation of information. However, the widespread sharing of inaccurate and misleading content has created significant challenges for information reliability and public trust. Consequently, researchers have developed various computational techniques to detect misinformation and assess the credibility of online content.

 This work is licensed under a Creative Commons Attribution 4.0 International License

2.2 Traditional Misinformation Detection Techniques

Traditional misinformation detection systems mainly focus on identifying linguistic patterns and statistical characteristics within textual information. These approaches analyse features such as word frequency, sentence structure, writing style, keywords, and sentiment to distinguish potentially misleading content from reliable information.

- Keyword-based identification.
- Linguistic feature analysis.
- Statistical text classification.
- Sentiment analysis.
- Source-based credibility assessment.

2.3 Machine Learning-Based Misinformation Detection

Machine learning has significantly improved the automation of misinformation detection by enabling classification models to learn patterns from labelled datasets.

The general process involves collecting data, preprocessing textual content, extracting relevant features, training classification algorithms, and evaluating the resulting models.

Commonly used machine learning algorithms include

- Naive Bayes.
- Support Vector Machine (SVM).
- Random Forest.
- Logistic Regression.
- Decision Tree.

2.4 Multimodal Misinformation Detection

Modern misinformation is not restricted to textual content. It may involve misleading images, manipulated videos, audio recordings, and combinations of different information types.

1. **Textual Feature Extraction:** Identifies linguistic and semantic characteristics using NLP and transformer-based models.
2. **Visual Feature Extraction:** Analyses associated images to identify relevant visual information and potential inconsistencies.
3. **Contextual Feature Extraction:** Examines available source information and contextual characteristics associated with submitted content.
4. **Multimodal Feature Fusion:** Combines representations obtained from different modalities to develop an integrated representation.
5. **Classification:** Uses the integrated representation to classify submitted content.

III. RESEARCH METHODOLOGY

3.1 Introduction

The proposed research methodology presents an Explainable Artificial Intelligence (XAI)-based framework for detecting misinformation and assessing the credibility of digital content. The framework integrates Natural Language Processing (NLP), Machine Learning (ML), transformer-based language models, multimodal feature extraction, and explainability techniques.

3.2 Proposed System Architecture

The proposed framework consists of the following major components:

1. Input Data Collection
2. Data Preprocessing
3. Textual Feature Extraction
4. Visual Feature Extraction
5. Contextual Feature Extraction
6. Multimodal Feature Fusion
7. Misinformation Classification
8. Credibility Assessment
9. Explainable AI Integration
10. Result Visualization

The system accepts textual information and associated images or contextual information wherever available. The

extracted features are processed through a classification model to generate predictions. Explainability techniques are then applied to identify the major factors influencing the model's decisions.

Operational Workflow & Database Schema

Steps	Process	Description	Output
1	Input Collection	Accept text and associated images.	Input data
2	Preprocessing	Clean text and process images.	Processed data
3	Feature Extraction	Extract textual, visual and contextual features.	Feature vectors
4	Feature Fusion	Combine extracted multimodal features.	Combined features
5	Classification	Predict content category using ML/DL.	Classification
6	Credibility Assessment	Generate model-based credibility score.	Credibility score
7	Explainability	Apply LIME and SHAP.	Model explanations
8	Result Generation	Display classification and explanations.	Final report

IV. EVALUATION PLAN

The evaluation plan describes the procedures used to assess the performance, reliability, and effectiveness of the proposed Explainable AI (XAI)-based multimodal misinformation detection framework.

4.1 Evaluation Objectives

The main objectives of the evaluation are:

- To measure the accuracy of misinformation classification.
- To evaluate the performance of NLP and transformer-based models.
- To examine the contribution of textual, visual, and contextual features.
- To assess the reliability of credibility scores and confidence estimates.
- To evaluate the effectiveness of LIME and SHAP explanations.
- To analyses the overall performance of the proposed framework.

4.2 Performance Evaluation Metrics

Metric	Description
Accuracy	Measures the overall correctness of predictions.
Precision	Measures how many predicted misleading items are actually misleading.
Recall	Measures how many actual misleading items are correctly identified.
F1-Score	Represents the balance between precision and recall.
Confusion Matrix	Shows correct and incorrect classification results.

V. DISCUSSION AND ANALYSIS

1. **Practicality:** The proposed Explainable AI-based multimodal misinformation detection framework is designed to be practically applicable through an interactive web-based application.
The system can be developed using Python and existing machine learning, NLP, deep learning, and explainability libraries. A framework such as Streamlit can provide a simple interface for users to submit textual content and associated images.

2. Explainability: Explainability is a central feature of the proposed framework. Traditional AI classification systems may produce predictions without clearly identifying the factors influencing their decisions.

2. **Limitations:** The proposed Explainable AI-based multimodal misinformation detection framework has certain theoretical limitations that may affect its accuracy, reliability, and general applicability.

Data Flow & System Architecture Diagram

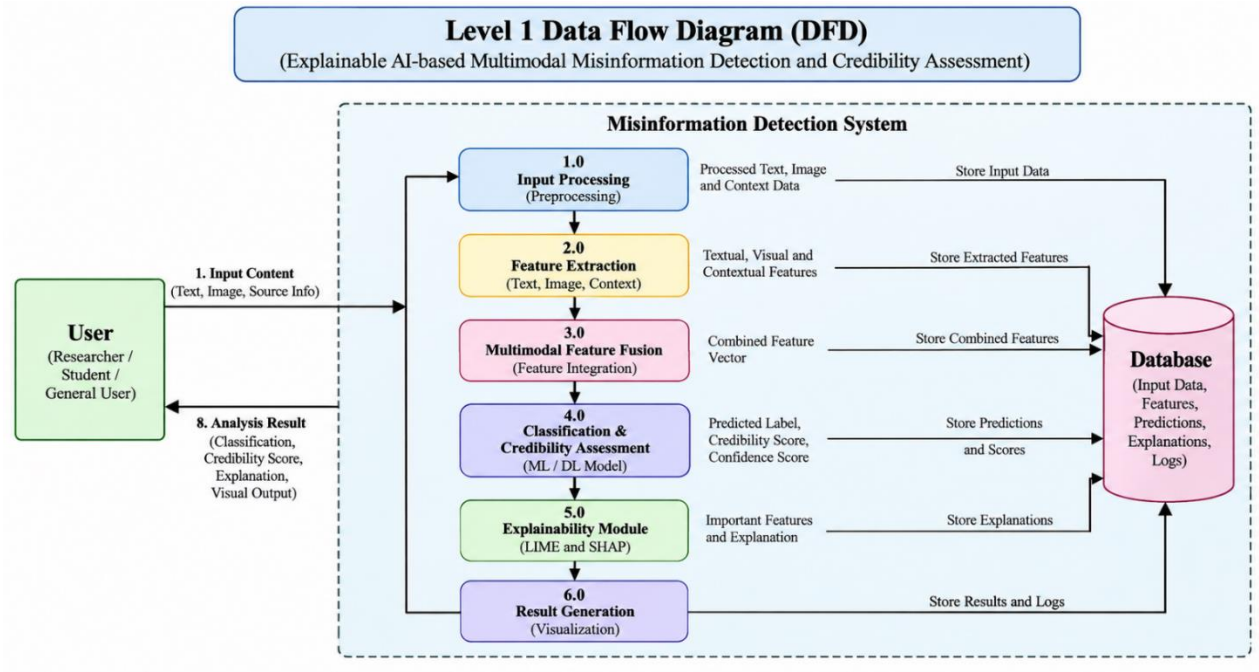


Fig. 1 Explainable AI-Based Multimodel Misinformation Detection and Credibility Assessment

VI. CONCLUSION

The proposed research presents an Explainable Artificial Intelligence (XAI)-based framework for multimodal misinformation detection and credibility assessment. The framework integrates Natural Language Processing, machine learning, transformer models, and multimodal feature extraction to analyse textual, visual, and contextual information. Overall, the proposed framework aims to improve transparency, reliability, and user understanding in misinformation detection. It provides a foundation for developing an intelligent and explainable system to support the assessment of digital.

REFERENCES

[1]. D. M. J. Lazer et al., "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018. DOI: https://doi.org/10.1126/science.aao2998.

[2]. H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017. DOI: https://doi.org/10.1257/jep.31.2.211.

[3]. 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. DOI: https://doi.org/10.1145/3137597.3137600.

[4]. X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, Article 109, 2020. DOI: https://doi.org/10.1145/3395046.

[5]. W. Y. Wang, "'Liar, Liar Pants on Fire': A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 422–426, 2017. DOI: https://doi.org/10.18653/v1/P17-2067.

[6]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. DOI: https://doi.org/10.18653/v1/N19-1423.

[7]. A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8]. Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 849–857, 2018. DOI: https://doi.org/10.1145/3219819.3219903.

[9]. K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6149–6157, 2020.

 This work is licensed under a Creative Commons Attribution 4.0 International License

[10]. K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 395–405, 2019. DOI: https://doi.org/10.1145/3292500.3330935.

[11]. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. DOI: https://doi.org/10.1145/2939672.2939778.

[12]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[13]. M. Szczepański, M. Pawlicki, R. Kozik, and M. Choraś, "New explainability method for BERT-based model in fake news detection," *Scientific Reports*, vol. 11, Article 23705, 2021. DOI: https://doi.org/10.1038/s41598-021-03100-6.

[14]. M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, and A. Vlachos, "Multimodal automated fact-checking: A survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5430–5448, 2023. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.361.

[15]. C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: A survey of deep learning techniques," *Social Network Analysis and Mining*, vol. 13, Article 101, 2023. DOI: https://doi.org/10.1007/s13278-023-01104-w.

[16]. J. Lv et al., "Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges," *Journal of King Saud University Computer and Information Sciences*, vol. 37, Article 306, 2025. DOI: https://doi.org/10.1007/s44443-025-00317-7.

[17]. K. Mishima and H. Yamana, "A survey on explainable fake news detection," *IEICE Transactions on Information and Systems*, vol. E105.D, no. 7, pp. 1249–1257, 2022. DOI: https://doi.org/10.1587/transinf.2021EDR0003.

[18]. A. B. Athira, S. D. Madhu Kumar, and A. M. Chacko, "A systematic survey on explainable AI applied to fake news detection," *Engineering Applications of Artificial Intelligence*, vol. 122, Article 106087, 2023. DOI: https://doi.org/10.1016/j.engappai.2023.106087