iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79064

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Bridging the Gap Between Multimodal Sentiment Analysis and Explainable AI: A Conceptual Framework

Akshatha Rithesh, Divya M O, Neethu K, Sreeshma Mohan
S-VYASA Deemed To Be University

Abstract: ***The rapid development of social media has raised the importance of sentiment analysis in the context of user emotions. Recent developments in multimodal sentiment analysis have incorporated various features such as text, emojis, and hashtags to improve the predictive accuracy of the system. However, most of the existing multimodal sentiment analysis models are complex and involve deep learning and language models that are black box in nature. This makes it difficult for the user to trust the system. On the other hand, Explainable Artificial Intelligence (XAI) has been developed to improve the transparency of the system. However, the integration of multimodal sentiment analysis and Explainable Artificial Intelligence is in its infancy.The current paper presents a conceptual framework that bridges the gap in multimodal sentiment analysis and Explainable Artificial Intelligence. The proposed framework incorporates various features such as text, emojis, and hashtags with the Explainable Artificial Intelligence approach. The proposed approach will improve the transparency and usability of the multimodal sentiment analysis system. The research challenges and implementation of the proposed approach are also highlighted in the current study.***

Keywords: Multimodal Sentiment Analysis, Explainable AI, Emojis, Hashtags, Interpretability, Human-Centric AI

I. INTRODUCTION

The role of sentiment analysis in understanding user opinions and emotions through digital media such as social media, online product reviews, and online forums is significant. With the growing volume of user-generated content, there is a need to use machine learning models to automate the process of sentiment analysis. Earlier approaches primarily focused on textual data; however, recent studies highlight the importance of incorporating additional contextual elements such as emojis, which carry rich emotional information [1].

Advancements in deep learning, particularly transformer-based models, have significantly improved the performance of sentiment analysis systems by capturing complex contextual relationships [2]. More recent developments have focused on multimodal sentiment analysis, which integrates multiple sources of information such as text, emojis, and hashtags to enhance prediction accuracy [3]–[5]. Despite these improvements, such models are often considered black-box systems, making it difficult to interpret their decisions [6].

Explainable Artificial Intelligence (XAI) has emerged as a critical research area to address this limitation by providing transparency and interpretability in AI systems [11]–[14]. In addition, recent studies have explored the integration of explainability into multimodal and advanced AI systems, highlighting the importance of trust and human-centric design [7]–[10], [15]. However, most existing approaches still treat multimodal sentiment analysis and explainable AI as separate domains.

Therefore, there remains a significant gap in developing a unified and human-centric framework that effectively integrates multimodal sentiment analysis with explainable AI.Assuch, this paper aims to propose a conceptual framework that integrates multimodal sentiment analysis and explainable AI using textual features, emoji features, and hashtag features.

II. LITERATURE REVIEW

Sentiment analysis has witnessed significant advancements with the evolution of deep learning techniques and the increasing availability of multimodal data. Early studies demonstrated that emojis carry rich emotional information and can enhance sentiment representation, marking a shift from traditional text-based approaches to more expressive sentiment analysis methods [1].

The introduction of transformer-based models further improved the performance of sentiment analysis systems by enabling deeper contextual understanding of textual data [2]. These models significantly enhance prediction accuracy; however, they often operate as complex black-box systems, limiting interpretability.

Recent research has increasingly focused on multimodal sentiment analysis by integrating multiple data sources such as text, emojis, and hashtags. Interpretable multimodal approaches using large-scale language models have been proposed to combine performance with interpretability [3]. Additionally, contrastive knowledge injection techniques have been introduced to improve multimodal feature representation [4]. Cross-lingual multimodal sentiment analysis models further extend these approaches by demonstrating scalability across diverse languages and datasets [5].

Alongside these developments, explainability has become an essential requirement in sentiment analysis systems. Explainable pre-trained language models have been developed to incorporate interpretability into modern architectures [6]. Survey studies further emphasize the importance of integrating explainable artificial intelligence into sentiment analysis to improve transparency and trust [7]. Moreover, causal reasoning-based approaches have been proposed to enhance interpretability by identifying relationships between input features and predictions [8].

Recent research also highlights broader trends and applications in sentiment analysis. Analytical studies indicate the growing role of deep learning and multimodal approaches in advancing sentiment analysis techniques [9]. Furthermore, the integration of explainable AI into multimodal sentiment systems has been explored to balance predictive performance with interpretability in real-world applications [10].

The foundation of explainable artificial intelligence is built upon model-agnostic techniques that provide insights into model decisions. Local explanation methods and feature attribution techniques have been widely adopted to interpret machine learning models [11], [12]. Subsequent research has emphasized the importance of transparency, trust, and accountability in AI systems [13], [14]. More recent studies highlight emerging challenges in explainable AI, including the need for human-centric explanations and improved evaluation mechanisms [15].

Despite these advancements, multimodal sentiment analysis and explainable artificial intelligence are largely explored as separate research domains. While multimodal approaches enhance prediction performance by leveraging diverse data sources, explainable AI techniques primarily focus on interpretability, often limited to text-based systems. This indicates a clear research gap in developing unified frameworks that integrate multimodal inputs with explainable and human-centric decision-making mechanisms.

A. Framework Overview

Below is a simple representation of the framework

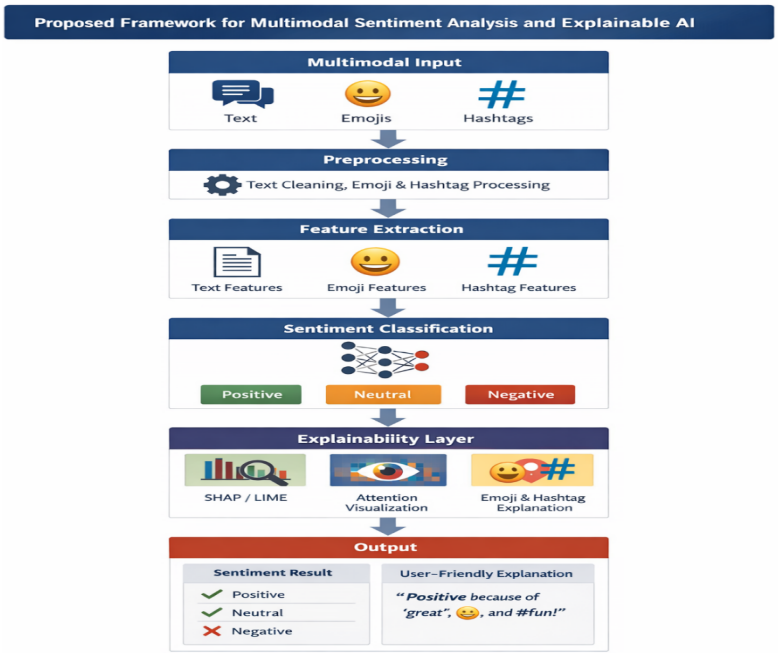


Fig. 1 Proposed model for Multimodal Sentiment Analysis and Explainable AI

B. Framework Components

1) Multimodal Input Layer

This layer gathers data from social media platforms, including textual content, emojis, and hashtags. Unlike traditional sentiment analysis approaches that primarily rely on textual information [2], the proposed framework incorporates multiple modalities to capture a richer and more nuanced representation of user emotions. The integration of these diverse inputs enables better contextual understanding and improves sentiment interpretation [3]–[5], [8].

2) Preprocessing Layer

In this stage, the collected data is cleaned and prepared for analysis. This includes text normalization, removal of noise and stop words, conversion of emojis into their corresponding meanings, and segmentation of hashtags into meaningful components. These steps ensure that all input modalities are structured and aligned for effective processing in subsequent stages.

3) Feature Extraction Layer

The feature extraction layer transforms the processed data into meaningful representations that can be used for sentiment prediction. Textual data is converted into embeddings using advanced deep learning models such as transformers [2]. Emojis are mapped to their corresponding emotional meanings to capture implicit sentiment [1], while hashtags are analyzed to extract contextual keywords. The combination of these features enhances the overall semantic understanding of the input data.

4) Sentiment Classification Model

In this stage, a machine learning or deep learning model is used to classify the sentiment into categories such as positive, negative, or neutral. Although modern models, particularly transformer-based architectures, achieve high accuracy, they often function as black-box systems, making it difficult to understand how decisions are made [6], [10].

5) Explainability Layer

This layer represents the key contribution of the proposed framework. It introduces explainability mechanisms to interpret model predictions and address the limitations of black-box models. Techniques such as feature importance methods (e.g., SHAP and LIME) are used to identify the contribution of different input features [11], [12]. In addition, attention visualization and rule-based explanations help provide deeper insights into the model's decision-making process. This layer ensures that the system generates transparent and trustworthy outputs, aligning with recent research emphasizing the importance of explainable AI [13]–[15].

6) Output Layer

The final output includes:

- Sentiment classification (Positive/Negative/Neutral)
- Human-understandable explanation

Example:

"The sentiment is Positive because of words 'happy', emoji , and hashtag #blessed."

C. Key Novelty of the Framework

The proposed framework introduces the following contributions:

- Integration of multimodal inputs (text + emojis + hashtags)
- Incorporation of a dedicated explainability layer
- Generation of human-centric explanations
- Bridging the gap between performance and interpretability

Unlike existing approaches that treat multimodal learning and XAI separately, this framework unifies both into a single pipeline.

D. Practical Significance

The proposed framework can be applied in:

- Mental health monitoring
- Social media analytics
- Customer feedback systems

By providing both predictions and explanations, the system enhances trust, usability, and decision-making.

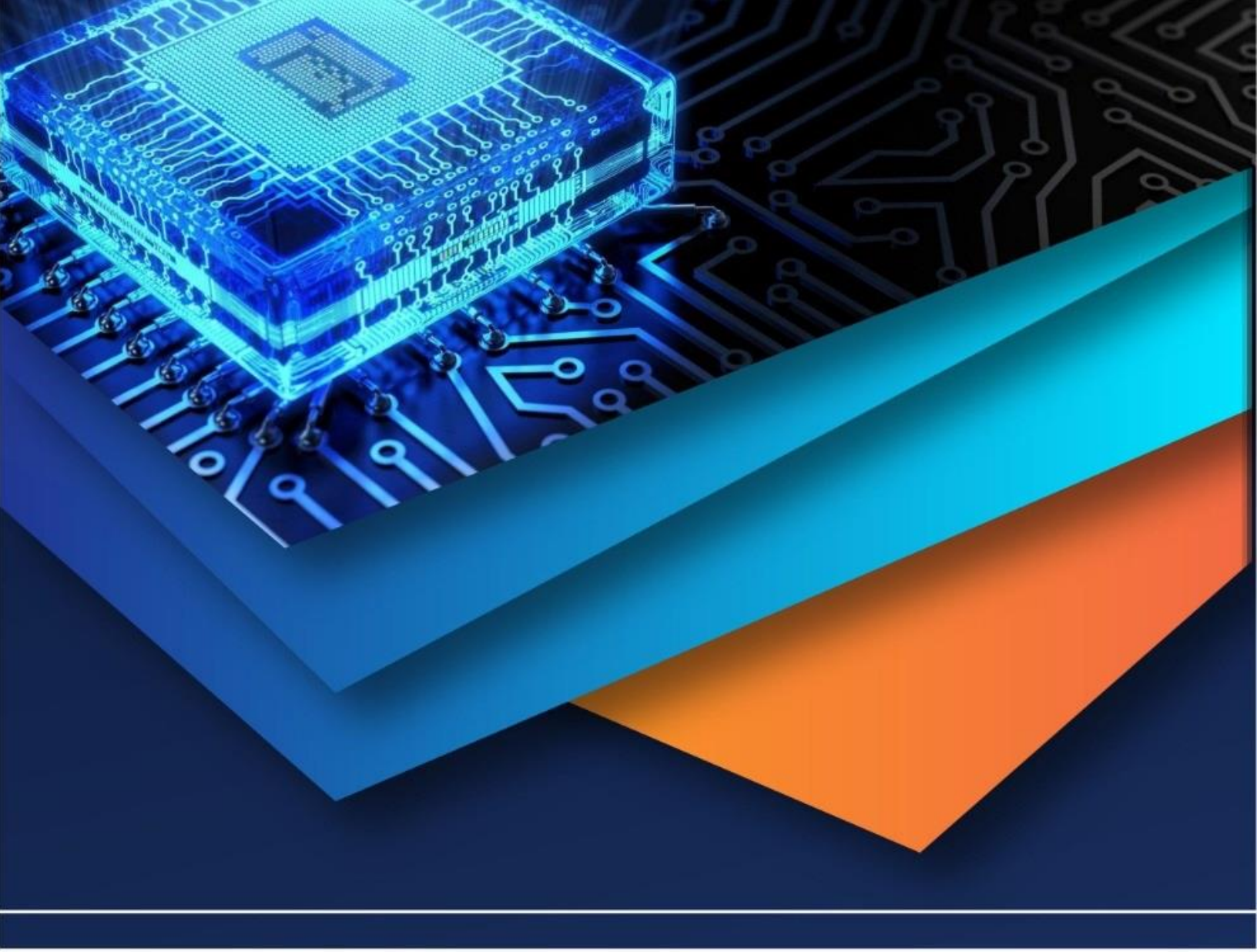
III. CONCLUSION AND FUTURE WORK

This paper presented a conceptual framework aimed at bridging the gap between multimodal sentiment analysis and explainable artificial intelligence. Although recent approaches have significantly improved sentiment prediction accuracy through deep learning and multimodal integration, they often operate as complex black-box systems, limiting transparency and interpretability [2]–[5], [8]–[10]. At the same time, explainable AI techniques have made considerable progress in improving the understanding of model decisions; however, these methods are largely focused on text-based systems and do not fully address multimodal scenarios [11]–[14].

To overcome these limitations, the proposed framework integrates textual data, emojis, and hashtags with explainability mechanisms to provide both accurate and interpretable sentiment predictions. By introducing a dedicated explainability layer, the framework enables the identification of feature contributions and generates human-understandable explanations, thereby enhancing user trust and system usability. This unified approach contributes to the development of transparent and human-centric sentiment analysis systems, in line with recent advancements in explainable and responsible AI [6], [15].

REFERENCES

[1] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).
[3] Li, S., & Okada, S. (2023). Interpretable multimodal sentiment analysis based on textual modality descriptions using large-scale language models. arXiv preprint.
[4] Yu, Y., Zhao, M., Qi, S., Sun, F., Wang, B., Guo, W., Wang, X., Yang, L., & Niu, D. (2023). ConKI: Contrastive knowledge injection for multimodal sentiment analysis. arXiv preprint.
[5] Miah, M. S. U., et al. (2024). A multimodal approach to cross-lingual sentiment analysis using transformers and large language models. Scientific Reports.
[6] Mabokela, K. R., et al. (2024). Explainable pre-trained language models for sentiment analysis. Big Data and Cognitive Computing.
[7] Diwali, A. (2024). Sentiment analysis meets explainable artificial intelligence: A survey. IEEE Transactions.
[8] Chen, F., Huang, P., Ge, X., Huang, J., & Bao, Z. (2024). Multimodal sentiment analysis based on causal reasoning. arXiv preprint.
[9] Hill, C. (2025). An analytical assessment of sentiment analysis trends and applications (2012–2024). ScienceDirect.
[10] Dolhopolov, S., Riabchun, Y., Delembovskyi, M., &Molodid, O. (2026). Explainable artificial intelligence for multimodal sentiment analysis in revitalization project management.
[11] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).
[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS).
[13] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access.
[14] Arrieta, A. B., Díaz-Rodríguez, N., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion.
[15] Longo, L. (2024). Explainable artificial intelligence (XAI) 2.0: Open challenges and future directions. Information Fusion.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY