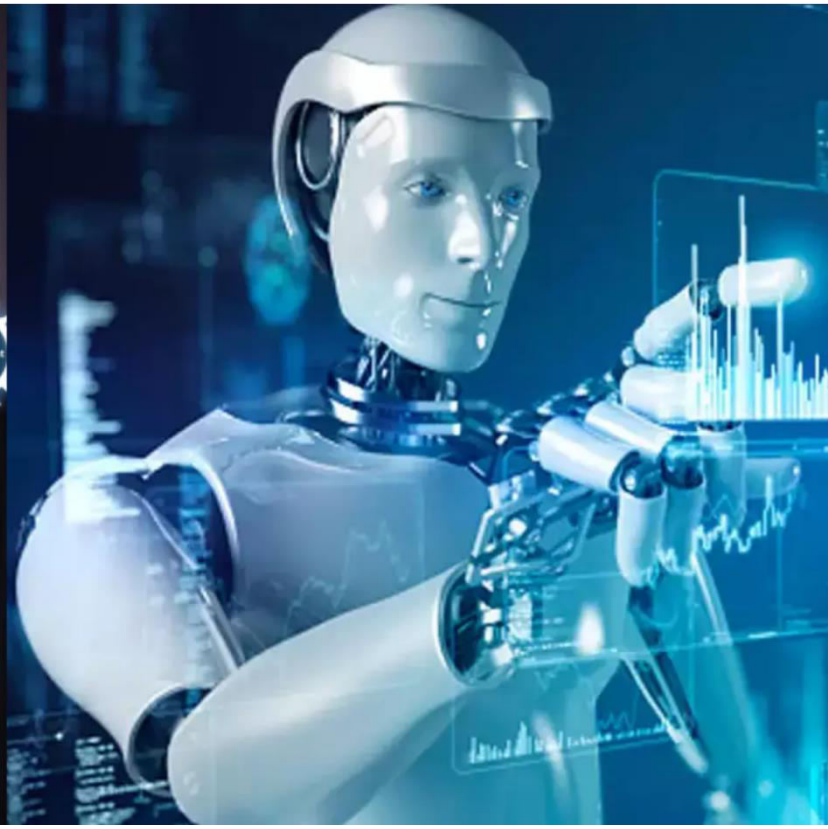
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503169|

Multimodal Sentimental Analysis Based on Deep Learning: A Review

Rutuja Madhav Gangavane

Department of Computer Engineering, Ajeenkya DY Patil School of Engineering, Pune, India

ABSTRACT: This study presents a deep-learning-based multimodal sentiment analysis system that integrates textual, visual, and audio information to improve emotion understanding. Traditional text-only models often fail to capture hidden cues such as tone, facial expressions, and visual context, leading to inaccurate predictions. The proposed system uses separate feature extraction modules for each modality and a fusion mechanism that combines their representations using attention-based deep learning. This unified feature vector is then classified to determine sentiment polarity. The system enhances accuracy, reduces noise in unimodal data, and performs efficiently on complex real-world inputs such as YouTube reviews and social media posts. Overall, this multimodal framework provides a more comprehensive and reliable sentiment interpretation than conventional single-modality approaches.

KEYWORDS: Multimodal Sentiment Analysis, Deep Learning, Feature Fusion, Attention Mechanism, Emotion Detection, Text-Audio-Video Processing

I. INTRODUCTION

Sentiment analysis plays a major role in understanding human emotions expressed across digital platforms. With the growth of social media, users share opinions through short texts, images, and spoken videos, making emotion detection more complex. Traditional text-based models often struggle with sarcasm, ambiguous language, and missing emotional cues. To overcome these challenges, multimodal analysis combines multiple data sources for a deeper understanding. By integrating text, audio, and visual signals, the system captures complete emotional context. This approach significantly improves accuracy compared to unimodal methods.

Recent advances in deep learning have enabled automatic extraction of meaningful features from raw multimedia input. Models such as CNNs, RNNs, and Transformers can identify patterns that were previously difficult to analyze manually. Visual cues like facial expressions, audio cues such as pitch and tone, and textual semantics work together to reveal the true sentiment. Multimodal fusion helps the system minimize noise from any single modality. This holistic view supports more reliable predictions even when one data type is unclear or incomplete. Thus, multimodal sentiment analysis has become essential in modern applications.

As digital content continues to expand, the demand for accurate sentiment understanding grows rapidly. Businesses require emotion insights for customer feedback, governments need it for public opinion monitoring, and mental-health systems use it for early detection of distress. However, challenges remain in handling cross-modal alignment, feature fusion, and varying content quality. Deep learning helps address these issues by learning high-level representations automatically. The proposed system aims to enhance fusion techniques and improve prediction accuracy. This research contributes to building more intelligent, human-aware systems.

II. MOTIVATION

The primary motivation is to overcome the limitations of text-only sentiment analysis. Emotions are often expressed through voice modulation, facial expressions, and visual context, which cannot be captured by text alone. Multimodal systems help decode hidden emotions, reduce misclassification, and improve decision-making in real-world applications such as social media monitoring, customer service, and mental-health support.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503169|

III. LITERATURE SURVEY

Jianfei Yu [1], we study Target-oriented Multimodal Sentiment Classification (TMSC) and propose a multimodal BERT architecture. To model intra-modality dynamics, we first apply BERT to obtain target-sensitive textual representations. We then borrow the idea from selfattention and design a target attention mechanism to perform target-image matching to derive target sensitive visual representations.

Bing Liu [2], Deep learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results. Along with the success of deep learning in many other application domains, deep learning is also popularly used in sentiment analysis in recent years. This paper first gives an overview of deep learning and then provides a comprehensive survey of its current applications in sentiment analysis.

Yao-Hung Hubert Tsai [3], At the heart of our model is the directional pairwise cross modal attention, which attends to interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another. Comprehensive experiments on both aligned and non-aligned multi modal time-series show that our model outper forms state-of-the-art methods by a large mar gin. In addition, empirical analysis suggests that correlated crossmodal signals are able to be captured by the proposed crossmodalattention mechanism in MulT.

Nan Xu [4], In the context of aspect level sentiment analysis, multimodal data are often more important than text-only data, and have various correlations in cluding impacts that aspect brings to text and image as well as the interactions associated with text and image. However, there has not been any related work carried out so far at the intersection of aspect-level and multimodal sentiment analy sis. To fill this gap, we are among the first to put forward the new task, aspect based multimodal sentiment analysis, and propose a novel Multi-Interactive Memory Network (MIMN) model for this task.

AnirudhBindiganavale Harish [5], In contrast, we propose an attention-based deep neural network and a training approach to facilitate both feature and decision level fusion. Our network effectively lever ages information across all three modalities using a 2 stage fusion process. We test our network on the individual utter ance based contextual information extracted from the CMU MOSI Dataset. A comparison is drawn between the state-of the-art and our network.

Lujuan Deng [6], this paper proposesamodelbasedonamulti-headattentionmechanism.First,afterpre-processingtheoriginaldata.Then,the feature representation is converted into a sequence of word vectors and positional encoding is introduced to better understand the semantic and sequential information in the input sequence. Next, the input coding sequence is fed into the transformer model for further processing and learning. At the transformer layer, a cross-modal attention consisting of a pair of multi-head attention modules is employed to reflect the correlation between modalities.

IV. GAP ANALYSIS

Existing models focus mainly on unimodal or partially multimodal approaches, leading to loss of contextual information. Many systems lack efficient fusion strategies, resulting in poor integration of text, audio, and visual features. Traditional machine learning requires extensive manual feature engineering. Moreover, current models struggle with unaligned multimodal data, complex backgrounds, and sentiment ambiguity.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503169|

V. SYSTEM ARCHITECTURE

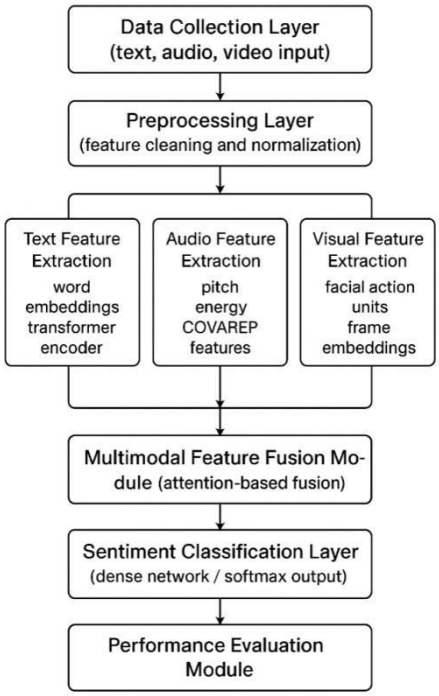


Figure 1. System Architecture

VI. PROPOSED SYSTEM

The proposed system introduces an end-to-end deep learning framework that processes text, audio, and visual data separately and then merges them using an attention-based fusion mechanism. Text features are encoded using transformer embeddings, while audio features are extracted through COVAREP-based acoustic analysis. Visual cues are obtained from facial expression embeddings. These features are fused into a shared representation, allowing the classifier to understand emotional patterns from multiple perspectives. This multimodal integration results in improved robustness and sentiment prediction accuracy compared to traditional methods.

VII. CONCLUSION

Multimodal sentiment analysis represents a major advancement in emotion detection by combining text, audio, and visual information. The proposed deep learning framework successfully enhances feature extraction, improves fusion strategies, and increases prediction accuracy. By leveraging transformers and attention mechanisms, the system captures richer emotional cues and overcomes limitations of unimodal approaches. This results in more reliable and human-like sentiment interpretation.

VIII. FUTURE SCOPE

- Integrating real-time multimodal emotion monitoring systems
- Expanding datasets to include multilingual and cross-cultural sentiment variations
- Developing lightweight models for mobile and IoT devices
- Enhancing noise-handling in low-quality videos or audio
- Applying multimodal emotion analysis to healthcare, customer experience, and virtual assistants

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503169|

REFERENCES

[1] P. Lin, X. Luo and Y. Fan: "A Survey of Sentiment Analysis Based on Deep Learning", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering,2020, Vol.14, No.12, pp.473-485

[2] Harish A. and Sadat F., "Trimodal attention module for multimodal sentiment analysis," in Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020,pp. 13803-13804.

[3] X. Chen, G.-M. Lu and J. Yan, and Yan J., "Multimodal sentiment analysis based on multi-head attention mechanism", Proceedings of the 4th International Conference on Machine Learning and Soft Computing, 2020, pp. 34–39.

[4] J.-F. Yu and J. Jiang, "Adapting bert for targetoriented multimodal sentiment classification", Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, pp. 5408–5414.

[5] N. Xu, W. Mao and G. Chen, "MultiInteractive Memory Network for Aspect Based Multimodal Sentiment Analysis", Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, No. 1, pp.371-378.

[6] Tsai Yao-Hung Hubert, Bai Shaojie, Pu Liang Paul, Kolter J Zico, Morency Louis-Philippe, SalakhutdinovRuslan., "Multimodal Transformer for Unaligned Multimodal Language Sequences", Proceedings of the conference Association for Computational Linguistics, 2019, pp.6558-6569.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details