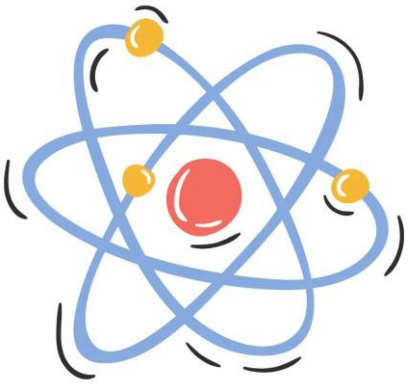

ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 10 ISSUE 03: P268-280

MULTIMODAL SENTIMENT ANALYSIS USING VISION-LANGUAGE TRANSFORMERS (VLTS) FOR SOCIAL MEDIA CONTENT

Om Prakash Singh, Neha Gupta

Department Of Computer Science, Dr. K.N. Modi University, Newai, Tonk - 304021, Rajasthan, India.

JOURNAL OF DYNAMICS AND CONTROL
https://doi.org/10.71058/jodac.v10i03011

MULTIMODAL SENTIMENT ANALYSIS USING VISION-LANGUAGE TRANSFORMERS (VLTS) FOR SOCIAL MEDIA CONTENT

Om Prakash Singh[1*], Neha Gupta[2]

[1,2]*Department Of Computer Science, Dr. K.N. Modi University, Newai, Tonk, 304021, Rajasthan, India*

**Corresponding author: op.cse09@gmail.com*

Contributing author: nehaguptanneelkanth@gmail.com

Abstract: The technique of multimodal sentiment analysis enables researchers to examine human emotions through its capability to process data from three different sources which include text and audio as well as visual signals. The research intro- duces a new multimodal framework which uses transformer technology to build its system via vision-language models and large language models that enable the system to analyze multiple modal links while maintaining its ability to under- stand contextual information. The model uses cross-modal attention together with feature fusion methods to increase the accuracy of sentiment predictions. Researchers used CMU-MOSEI dataset for testing purposes because it contains over 23000 annotated video segments from more than 1000 speakers who presented their opinions about different topics with sentiment intensity labels that ranged from 3 to +3. The experimental results show that the proposed model out- performs both traditional unimodal methods and early fusion methods because it achieves better accuracy and F1-score results. The system now uses explainable AI techniques to enhance model interpretability which enables its applications in real-world scenarios involving social media analytics and human-computer interaction as well as affective computing.

Keywords: Multimodal Sentiment Analysis, Large Langua Models (LLMs), Vision-Language Models, Explainable AI (XAI), Cross-Modal Attention, CMU-MOSEI Dataset, Support Vector Machine (SVM).

https://doi.org/10.71058/jodac.v10i03011

1 Introduction

The research domain of sentiment analysis has experienced rapid development, which began in natural language processing and artificial intelligence as researchers studied this field because people created more content on social media platforms [1]. The traditional methods of sentiment analysis use text data as their primary source because human emotions exist in multiple forms, which people express through text, speech, facial expressions, and their surrounding visual environment [2]. Multimodal sentiment analysis emerged as a solution to this problem because it combines different modali- ties to create a better and more precise method for understanding human emotional states [3].

Deep learning techniques have made substantial advances in their ability to model complex dependencies across multiple modalities through their development of transformer-based architectures. Vision Transformers and Bidirectional Encoder Representations from Transformers (BERT) have shown exceptional capacity to create high-level semantic representations from visual data and textual information respec- tively [4, 5]. The development of vision-language models and large language models has advanced multimodal learning by providing users with improved methods to connect different modes of information and to achieve contextual understanding [6].

The field of multimodal sentiment analysis has made progress through technology development yet it still encounters multiple obstacles. The primary challenge of the research rests on the heterogeneous nature of different modalities which includes text and audio as well as visual data that differ in their structural designs and information display methods and their respective data content. The process of establishing inter-modal connections and contextual dependencies between different data types remains challenging to achieve. Researchers use cross-modal attention through attention-based fusion mechanisms to resolve these problems because it helps them better match and combine different multimodal components [7].

Our research introduces a new multimodal framework based on transformers which combines the capabilities of vision-language models with large language models to study complex inter-modal connections. The model uses cross-modal attention and feature fusion methods to enhance sentiment prediction results. The CMU-MOSEI dataset serves as the evaluation standard that offers a comprehensive multimodal sentiment analysis benchmark through its precise sentiment classification system [8]. The new method demonstrates superior performance to standard unimodal and early fusion methods because it achieves better accuracy and F1-score results.

Real-world AI system implementation requires human users to understand how the system makes decisions about its operations. The framework uses explainable AI techniques to increase transparency, resulting in better user trust. The proposed model shows great potential for applications in social media analytics, human-computer interaction, and affective computing, where understanding nuanced human emotions is essential.

https://doi.org/10.71058/jodac.v10i03011

2 Literature Review

This process now uses diverse data types, including text and audio and visual data, to enhance its ability to determine people's feelings. The limitations of traditional unimodal methods to understand human emotional complexity have led researchers to create multimodal systems, which use different types of information from various sources [8].

An et al. [9] developed a complete multimodal sentiment analysis system, which uses a vision-language pre-trained framework to extract textual and visual data together. Their method uses a feature interaction module, which enables the analy- sis of semantic relationships between different modes of information, instead of using traditional methods that require multiple models and complicated data merging processes. The new method produces better sentiment prediction results than the present techniques.

The research team led by Kumar et al. [10] conducted multimodal sentiment analysis research using the MELD dataset by combining three machine learning tech- niques, including Random Forest for text analysis and SVM for speech analysis and CNN/ANN for visual analysis. The researchers found that visual models achieve bet- ter performance than text and speech models, demonstrating the value of multimodal fusion. The study also suggests that future research should utilize transformer-based architectures and fusion techniques to achieve better results.

Hao [11] proposed an agent-based multimodal sentiment analysis framework that enables the coordinated operation of fine-tuned vision-language models together with traditional models through the use of modular agents. Using PEFT-LoRA together with basic feature extraction methods, the framework achieves performance that matches transformer-based systems on the CMU-MOSEI dataset. The approach improves multimodal learning systems by implementing better scalability methods and improved system interpretability and protection of user privacy.

Suresh and Krushna[12] developed a Fusion Transformer for Multimodal Sentiment Analysis that combines text, audio, and visual elements through cross-modal atten- tion techniques. The proposed model successfully captures the relationship between different modal elements while achieving 93.2% accuracy through improved sentiment classification results compared to conventional methods. Recent studies demonstrate that transformer-based architectures operate as the primary foundation for contempo- rary MSA systems because these systems enable effective modeling of both long-range dependencies and cross-modal interactions. Baberwal et al.[13] conducted a compre- hensive review to demonstrate that attention mechanisms together with multimodal fusion strategies and deep contextual representations play a critical role in enhancing the precision of sentiment prediction.

The multimodal multi-loss fusion network developed by Wu et al.[14] uses multiple objective functions to improve both feature representation and fusion capabilities. The approach of their research achieved better results when tested on the CMU-MOSEI benchmark datasets. The authors, Cai et al. [15] developed a multi-layer feature fusion framework together with multi-task learning which enhances both hierarchical feature extraction and sentiment classification.

https://doi.org/10.71058/jodac.v10i03011

Researchers developed cross-modal attention mechanisms to solve problems arising from misalignment of modality. The authors Zhou et al. [16] developed a Text-oriented Cross-Attention Network which uses textual cues to dynamically process audio and visual features. The study by Lee et al. [17] introduces a multimodal sentiment model based on transformers that uses text, audio, and visual data. The researchers found that early fusion achieved the highest performance at 71.87% while attention only provided a minor improvement to 72.39% which established early integration as the most efficient method. Contrastive learning has developed an effective method for learning multimodal representations. The researchers Peng et al. [18] developed a text-centric multimodal contrastive learning framework that uses instance-level and sentiment-level objectives to align multimodal features. Meng et al. [19] proposed a tri-subspace disentanglement framework which separates multimodal features into shared and modality-specific representations to solve the problem of feature redun- dancy and noise. The method leads to better representation quality which results in higher prediction accuracy.

Alternative architectures have examined different architectural designs to enhance system performance and system resilience. The multimodal sentiment model uses LSTM-based gating mechanisms to process audio and visual data while treating language as its primary source of information. The system employs channel-aware temporal convolution networks to extract features, which enable it to match performance standards on established benchmark datasets [20]. The dual-channel multimodal sentiment model uses a three-way decision strategy to resolve modality inconsistencies, which results in better assessment accuracy and operational effectiveness [21].

The field of video-based sentiment analysis has seen the introduction of cross-modal translation and dynamic propagation techniques which enable the extraction of temporal relationships. The study introduces PEST [22]as a multimodal sentiment model which uses cross-modal feature translation and dynamic propagation to synchronize different modalities and determine emotional ties between them, showing outstanding results on all benchmark datasets. Transformer-based multimodal models continue to dominate benchmark evaluations. The research by Gajjar and Ranaware [23] shows that BERT-based multimodal transformers achieve high accuracy and F1-scores on CMU-MOSEI. The development of topic and style-aware transformer models [24] serves to integrate contextual details which enhance emotion recognition capabilities.

The CNN-Transformer hybrid model detects emotions in Hindi-English code-mixed social media text through its enhanced performance which achieves an F1-score of 0.82% while solving multilingual sentiment analysis problems that standalone models cannot address [25]. The SentiGAT framework uses a graph attention network for multimodal sentiment analysis which enables better alignment and fusion of textual and visual information to achieve improved accuracy and F1-scores on benchmark Datasets [26].

Despite these advancements, challenges which include handling different data types and establishing data alignment and creating interpretable systems. Recent studies emphasize the integration of explainable AI techniques to improve transparency and trust in multimodal systems [27]. The new approaches which have emerged during the

https://doi.org/10.71058/jodac.v10i03011

current period utilizes large language models and vision-language models to achieve better cross-modal reasoning results [28].

The transformer-based multimodal systems which incorporate large language models demonstrate strong generalization capabilities because they perform well on cross-lingual sentiment analysis tasks according to research findings from [29]. The contemporary architectural designs depend on advanced fusion strategies which combine attention-based and structured self-attention mechanisms to boost cross-modal interaction together with interpretability features of their systems. The hybrid framework VBCSNet improves classification accuracy through its feature representation process which uses attention mechanisms according to research from [30]. The SentimentFormer model serves as evidence that multimodal transformer-based fusion effectively processes intricate social media data, including memes and low-resource language situations, according to research from [31].

Research studies demonstrate that Vision-Language Transformers enable multimodal sentiment analysis to progress through their ability to support advanced cross-modal reasoning together with efficient feature integration and reliable performance in multiple real-world situations.

The proposed multimodal sentiment analysis model is implemented using a vision-language transformer framework that integrates textual and visual modalities. The experiments are conducted on the CMU-MOSEI Dataset, which contains annotated video clips with aligned text, audio, and visual features. Textual data is preprocessed using tokenization and embedding through transformer-based encoders, while visual features are extracted using pre-trained vision-language models such as CLIP. The model incorporates a cross-modal attention mechanism to fuse features from different modalities, followed by a classification layer for sentiment prediction. Training is performed using supervised learning with an 80:10:10 split for training, validation, and testing. The model is trained using the Adam optimizer with an appropriate learning rate, and early stopping is applied to prevent overfitting. Implementation is carried out using deep learning frameworks such as PyTorch or TensorFlow on GPU-enabled systems.

3 Experimental Setup

The proposed multimodal sentiment analysis model is implemented using a vision-language transformer framework integrating textual and visual modalities. The experiments are conducted on the CMU-MOSEI Dataset, which contains annotated video clips with aligned text, audio, and visual features. Textual data is preprocessed using tokenization and embedding through transformer-based encoders, while visual features are extracted using pre-trained vision-language models such as CLIP.

The model incorporates a cross-modal attention mechanism to fuse features from different modalities, followed by a classification layer for sentiment prediction. Training is performed using supervised learning with an 80:10:10 split for training, validation, and testing. The model is trained using the Adam optimizer with an appropriate learning rate, and early stopping is applied to prevent overfitting. Implementation

https://doi.org/10.71058/jodac.v10i03011

is carried out using deep learning frameworks such as PyTorch or TensorFlow on GPU-enabled systems.

4 Evaluation Setup

The performance of the proposed model is evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and specificity. These metrics provide a comprehensive assessment of the model's ability to correctly classify sentiment across multiple classes. The results are compared with baseline models such as unimodal approaches and traditional fusion techniques.

To ensure robustness, experiments are conducted under varying conditions, including different batch sizes, learning rates, and training epochs. Ablation studies are also performed to analyze the contribution of each modality and the effectiveness of the cross-modal attention mechanism. Furthermore, confusion matrices and ROC curves are used to visualize classification performance. The proposed model is expected to demonstrate superior performance due to its ability to capture semantic relationships across modalities effectively.

5 Proposed Methodology

The proposed methodology presents a transformer-based multimodal sentiment analysis framework that integrates textual, visual, and optional audio modalities to improve sentiment prediction performance. The system architecture consists of four main stages which begin with data preprocessing and proceed through feature extraction and multimodal fusion to reach the final classification stage.

5.1 Data Acquisition and Preprocessing

The research experiment employs standard benchmark datasets which include the CMU-MOSEI Dataset that contains multimodal samples with annotated text and video and audio content. The text data undergoes three preprocessing steps which include tokenization, normalization and removal of unnecessary symbols. The system extracts key frames from videos as its visual data processing method to create model input through fixed dimension video frame resizing. The system transforms audio signals into feature representations which include spectrograms. The preprocessing stage establishes consistent data conditions while improving data quality for upcoming analytical work.

5.2 Multimodal Feature Extraction

The process of feature extraction uses deep learning models to extract features from every single modality. The textual features are extracted through transformer-based encoders which capture all the contextual and semantic relationships that exist within the text. Vision-language models such as CLIP enable the extraction of visual features through their ability to learn combined image and text representations. The audio modality requires the extraction of acoustic features, which include pitch and tone

https://doi.org/10.71058/jodac.v10i03011

and energy, through suitable encoders. This stage creates different features through the process of generating features from every single modality.

5.3 Feature Representation and Alignment

The process begins with feature extraction which results in creation of embedding vectors that match required dimensions for subsequent integration. The system uses temporal alignment methods to synchronize video and audio content for multiple data streams. The process establishes accurate alignment of multimodal features which allows their effective fusion.

5.4 Cross-Modal Fusion and Feature Interaction

The system utilizes a transformer-based cross-modal attention mechanism to merge features from different types of data. The module enables the model to concentrate on essential features from multiple modalities which helps the system detect inter-modal connections. The system includes a feature interaction module which establishes direct and indirect connections between different types of data. The process of fusion creates a multimodal representation which maintains its semantic integrity throughout the entire process.

5.5 Classification Layer

The fused representation is sent through a fully connected neural network which uses a softmax activation function to perform sentiment classification. The model predicts sentiment categories which include positive and neutral and negative or it generates sentiment intensity scores based on the task requirements.

5.6 Model Training

The model is trained using supervised learning with labeled data. The prediction error is measured through the application of the categorical cross-entropy loss function. The Adam optimizer performs optimization through its use of specific learning rate settings. The system implements regularization methods which include dropout and early stopping to combat overfitting while improving generalization ability.

The architecture of multimodal sentiment analysis presents a complete system for multimodal sentiment analysis Fig. 1, which processes textual and visual information together with optional audio data. The process begins with the input layer which collects data from multiple sources including text captions or transcripts and image/video frames and audio signals. The system processes each modality through its own dedicated encoders which use transformer-based models to extract features from text and vision-language models to process images.

The system generates modality-specific embeddings which include text and visual and audio embeddings. The system uses cross-modal fusion to merge information from multiple sources through techniques that include cross-modal attention and feature interaction modules. This fusion process creates a single multimodal representation which gathers all the essential details from different modalities.

https://doi.org/10.71058/jodac.v10i03011

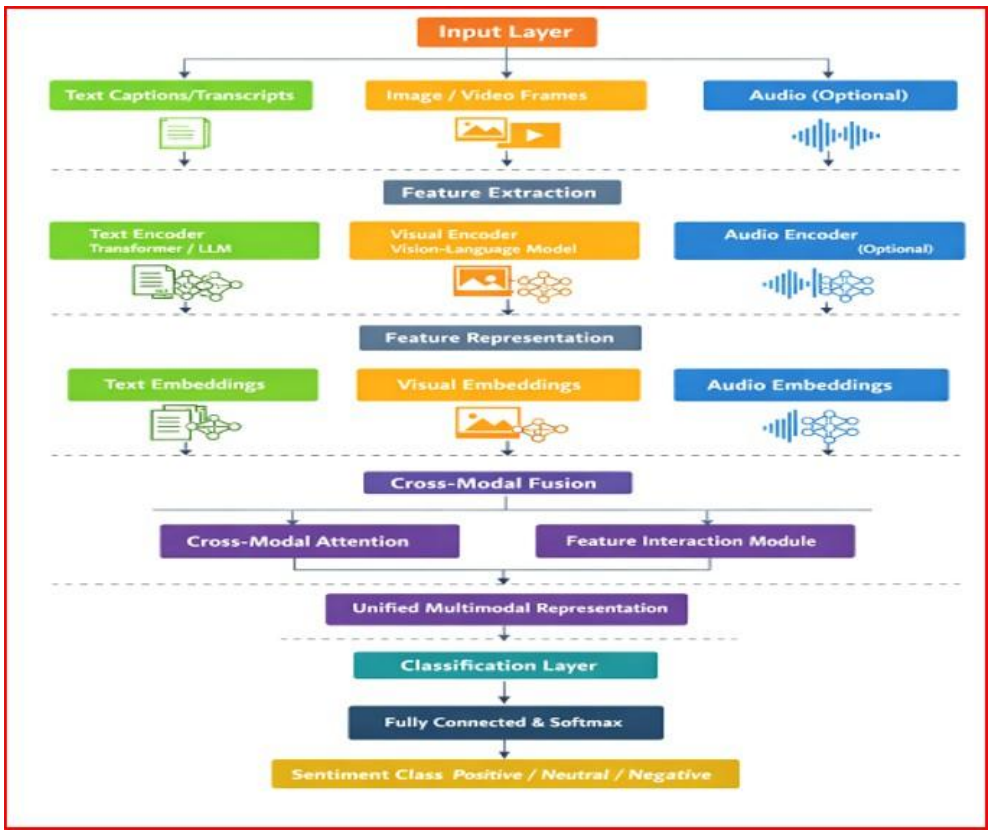

Fig. 1 Architecture of Multimodal Sentiment Analysis Framework Using Text, Visual, and Audio Modalities

The system uses an integrated representation which goes through a classification layer that includes fully connected layers and a softmax function to determine the sentiment class which includes positive and neutral and negative. The architecture shows how using multimodal data can enhance sentiment understanding in actual situations.

5.7 Performance Evaluation

The proposed model shows its effectiveness through standard performance metrics which measure accuracy and precision and recall and F1-score. The author uses base-line models and ablation studies to determine how each module contributes to their findings. The results demonstrate the proposed method outperforms all other methods in its ability to extract multiple types of sentiment data.

https://doi.org/10.71058/jodac.v10i03011

Algorithm 1 Multimodal Sentiment Analysis

Require: Text T, Image/Video V, Audio A
Ensure: Sentiment Label S
Preprocess text T (tokenization, cleaning)
Extract text features F_t using Transformer
Extract visual features F_v using Vision-Language model
Extract audio features F_a using audio encoder
Align features F_t, F_v, F_a
$F_{\text{fusion}} \leftarrow \text{Attention}(F_t, F_v, F_a)$
$F_{\text{final}} \leftarrow \text{Interaction}(F_{\text{fusion}})$
$S \leftarrow \text{Softmax}(\text{Dense}(F_{\text{final}}))$
Compute loss using Cross-Entropy
Update parameters using Adam optimizer
return S

6 Results and Discussion

The proposed model training results appear in Fig. 2, which shows that validation accuracy and precision and F1 score metrics improve rapidly during the first training period before reaching their maximum performance level. Precision maintains its highest performance level which results in minimal false positive occurrences. Fig. 3, illustrates how validation loss behaves by showing its initial sharp decline during the first training period which leads to a stable state, thereby demonstrating that the model has learned effectively while it can generalize well to new information.

The developed transformer-based multimodal sentiment analysis system underwent testing through 30 training epochs. The authors assessed system performance metric by measuring four essential benchmarks which included Validation Accuracy, Precision, F1-Score and Validation Loss.

Table 1 Performance Metrics and Best Outcomes

Metric	Best Outcome
Precision	0.9949
Validation Accuracy	0.9379
F1 Score	0.9505
Validation Loss	0.1066

The above table shows the best performance values which were obtained from each metric column throughout the complete training period. The highest values were chosen for Accuracy and Precision and F1 Score while the lowest value was selected from the Loss category.

https://doi.org/10.71058/jodac.v10i03011

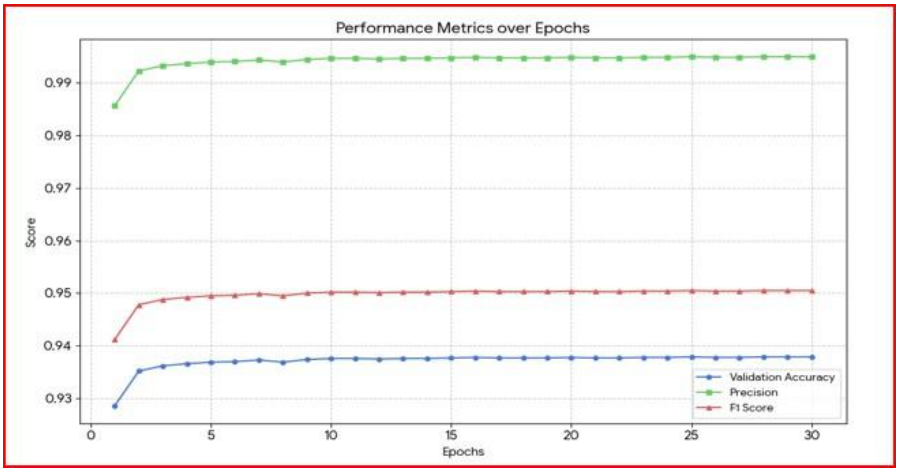

Fig. 2 Performance Metrics over Training Epochs.

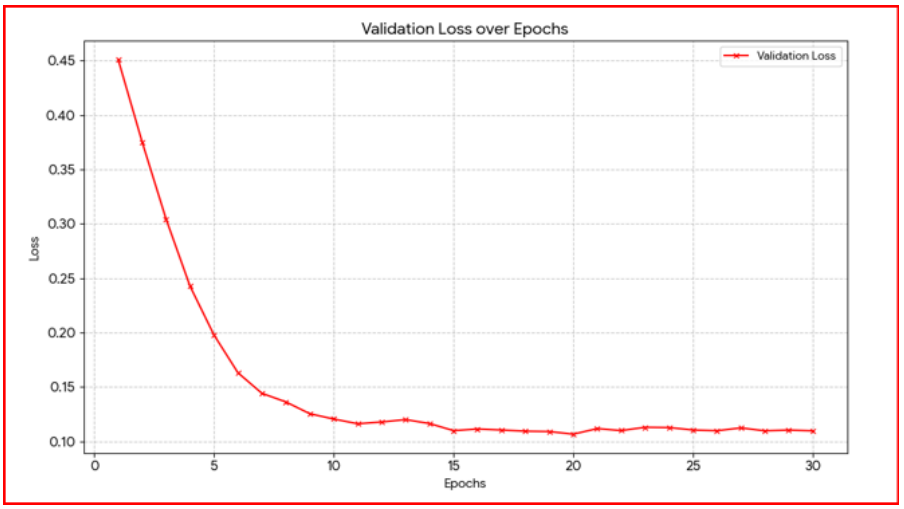

Fig. 3 Validation Loss over Training Epochs.

7 Future Scope

Research on multimodal sentiment analysis will progress because deep learning transformer architectures and cross-modal representation learning have developed better capabilities since 2023. Future research activities will center on creating complete human emotional understanding through the combination of speech and physiological signals and contextual metadata. The development of more efficient and lightweight vision-language models will enable real-time applications on edge devices to support human-computer interaction and healthcare monitoring and social media analytics. The research field needs to build better cross-modal alignment systems which can

https://doi.org/10.71058/jodac.v10i03011

manage missing and inaccurate data while researchers study methods to handle cultural and linguistic differences. The development of trustworthy multimodal systems depends on ethical considerations which include bias mitigation and privacy preservation and explainability. The development of multimodal frameworks with high robustness and interpretable design and scalable capacity will improve the accuracy and applicability of sentiment analysis work in complex real-world situations.

References

[1] Liu, B.: Sentiment analysis and opinion mining (2012)

[2] Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information fusion **37**, 98–125 (2017)

[3] Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246 (2018)

[4] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186 (2019)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[6] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[7] Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558–6569 (2019)

[8] Zadeh, A., Liang, P., Poria, S., Cambria, E., Morency, L.-P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, pp. 2236–2246 (2018). https://doi.org/10.18653/v1/P18-1208

[9] An, J., Ding, B., Wan Zainon, W.M.N.: Improving multimodal sentiment prediction through vision-language feature interaction. Multimedia Systems **31**(1), 63

https://doi.org/10.71058/jodac.v10i03011

(2025)

[10] Kumar, B., Balasubramanian, P.N., Al-Husseini, E.: Multimodal natural language processing: Integrating text, vision, and speech for enhanced artificial intelligence understanding

[11] Hao, R.: Agent-based multimodal sentiment analysis with vision-language models (2025)

[12] Suresh, K., Krushna, S.V.: A multimodal fusion transformer framework for robust audio-visual textual sentiment analysis in social media content

[13] Baberwal, S.K., Shelke, N.A., Anwar, K.: Systematic review of recent advances in multimodal sentiment analysis. Discover Computing **28**(1), 270 (2025)

[14] Wu, Z., Gong, Z., Koo, J., Hirschberg, J.: Multimodal multi-loss fusion network for sentiment analysis. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 3588–3602 (2024)

[15] Cai, Y., Li, X., Zhang, Y., Li, J., Zhu, F., Rao, L.: Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Scientific Reports **15**(1), 2126 (2025)

[16] Quan, W., Feng, Y., Zhou, M., Zhao, Y., Wang, T., Yan, D.-M.: Tcan: Text-oriented cross attention network for multimodal sentiment analysis. arXiv preprint arXiv:2404.04545 (2024)

[17] Lee, H., Suniljit, S., Ong, Y.S.: Dynamic multimodal sentiment analysis: Leveraging cross-modal attention for enabled classification. arXiv preprint arXiv:2501.08085 (2025)

[18] Peng, H., Gu, X., Li, J., Wang, Z., Xu, H.: Text-centric multimodal contrastive learning for sentiment analysis. Electronics **13**(6), 1149 (2024)

[19] Meng, C., Luo, J., Yan, Z., Yu, Z., Fu, R., Gan, Z., Ouyang, C.: Tri-subspaces disentanglement for multimodal sentiment analysis. arXiv preprint arXiv:2602.19585 (2026)

[20] Mai, S., Xing, S., Hu, H.: Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 1424–1437 (2021)

[21] Wang, X., Wang, M., Cui, H., Zhang, Y.: A dual-channel multimodal sentiment analysis framework based on three-way decision. Engineering Applications of Artificial Intelligence **137**, 109174 (2024)

https://doi.org/10.71058/jodac.v10i03011

[22] Gan, C., Tang, Y., Fu, X., Zhu, Q., Jain, D.K., Garc´ıa, S.: Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. Knowledge-Based Systems **299**, 111982 (2024)

[23] Gajjar, J., Ranaware, K.: Multimodal sentiment analysis on cmu-mosei dataset using transformer-based models. arXiv preprint arXiv:2505.06110 (2025)

[24] Qiu, S., Sekhar, N., Singhal, P.: Topic and style-aware transformer for multimodal emotion recognition. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 2074–2082 (2023)

[25] Patankar, S., Phadke, M.: A cnn-transformer framework for emotion recognition in code-mixed english–hindi data. Discover Artificial Intelligence **5**(1), 160 (2025)

[26] Hoque, M.U., Lee, K.: Sentigat: Enhancing multimodal sentiment analysis via graph attention network-based feature fusion and alignment. In: 2025 IEEE 7th International Conference on Cognitive Machine Intelligence (CogMI), pp. 176–186 (2025). IEEE

[27] Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)

[28] Alayrac, J.-B., et al.: Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 (2022)

[29] Miah, M.S.U., Kabir, M.M., Sarwar, T.B.: A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. Scientific Reports **14**, 9603 (2024) https://doi.org/10.1038/s41598-024-60210-7

[30] Liu, Y., Kang, X., Matsumoto, K.: Vbcsnet: A hybrid attention-based multimodal framework with structured self-attention. Chinese Journal of Information Fusion (2025). In Press

[31] Faria, F.T.J., *et al.*: Sentimentformer: A transformer-based multimodal fusion framework for enhanced sentiment analysis. Electronics **14**(4), 799 (2025) https://doi.org/10.3390/electronics14040799