International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Cross-Modal Sentiment Understanding: Fusing Text, Audio, and Vision for Deeper Emotion Analysis

Xiangshuai Huang
School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui Province, China, 233030
Email address: 2011330588@qqcom

***Abstract*—*With the rapid advancement of artificial intelligence, sentiment analysis has evolved into a crucial area of research in natural language processing and computer vision. While traditional approaches often rely on single modalities—such as text or speech—human emotions are inherently multimodal, conveyed through a combination of language, tone, facial expressions, and more. This paper focuses on cross-modal sentiment understanding, aiming to achieve deeper emotion recognition by integrating text, audio, and visual information. We recommend an innovative deep learning-driven framework that successfully integrates diverse modalities to improve sentiment classification. Experiments conducted on benchmark datasets demonstrate that our cross-modal fusion approach significantly outperforms unimodal baselines, highlighting the potential of joint multimodal representation learning in capturing complex emotional nuances.*

***Keywords*—*Multimodal Sentiment Analysis, Deep Learning, Text Analysis, Speech Analysis, Image Analysis, Sentiment Classification*.

I. INTRODUCTION

Understanding human emotions is central to building intelligent systems capable of natural and effective interaction. Sentiment analysis (SA), traditionally focused on text data, aims to identify emotions or opinions expressed through language. However, emotional expression in real-world communication extends far beyond text—it is inherently multimodal, involving vocal intonation, facial expressions, and body gestures that together convey affective states.

Relying solely on text overlooks critical non-verbal cues. For instance, the same sentence can express vastly different emotions depending on tone of voice, facial expressions, or visual context. This observation has led to the growing importance of Multimodal Sentiment Analysis (MSA), which integrates data from multiple modalities—such as text, audio, and vision—to achieve a richer and more accurate understanding of human emotion.

Recent progress in deep learning has significantly advanced cross-modal emotion understanding. Techniques such as Convolutional Neural Networks (CNNs) for extracting visual features, Recurrent Neural Networks (RNNs) for modeling sequential audio signals, and Transformer-based architectures for cross-modal fusion have enabled more nuanced sentiment recognition across modalities. These approaches leverage the complementary nature of different data sources: text provides semantic content, speech encodes emotional intensity and rhythm, and vision captures expressive facial dynamics.

This paper focuses on cross-modal sentiment understanding, exploring how the fusion of textual, auditory, and visual modalities can be orchestrated to improve sentiment prediction. We investigate multiple fusion strategies, feature extraction pipelines, and deep learning architectures that contribute to more robust and context-aware sentiment classification. By leveraging deep learning-based cross-modal integration, we aim to bridge the gap between isolated modality processing and holistic emotional comprehension. Methods for Multimodal Sentiment Analysis

II. METHODS FOR CROSS-MODAL SENTIMENT UNDERSTANDING

Multimodal sentiment analysis follows five main stages—Data Collection and Preprocessing, Modality-Specific Feature Extraction, Fusion of Features, Classification, and Evaluation. Each step is crucial for capturing and combining the diverse emotional signals present in text, speech, and vision.

A. Data Collection and Preprocessing

We select three benchmark corpora—CMU-MOSI, IEMOCAP and MELD—because each provides tightly aligned text, audio and video streams along with fine-grained emotion labels. CMU-MOSI consists of approximately 2,000 YouTube opinion clips rated on a continuous –3 to +3 polarity scale; IEMOCAP provides ten hours of scripted performances by actors categorized by primary emotions; and MELD contains over 1,400 natural conversational turns from *Friends*, annotated across seven categories from “anger” to “neutral.”

To prepare these multimodal streams for joint modeling, we first normalize and align timestamps across all three modalities. Text transcripts are lowercased, punctuation-stripped, tokenized (word- or subword-level) and matched to utterance timecodes. Audio tracks are then segmented into utterance clips using those same timecodes, denoised via a simple band-pass filter, amplitude-normalized, and stripped of leading/trailing silence through energy-threshold trimming. Video frames are sampled at a fixed rate (e.g. 5 fps), each frame is passed through a lightweight face detector (e.g. Haar or dlib), and detected faces are cropped and resized to a consistent resolution (e.g. 224×224). All three streams remain perfectly synchronized so that every text token, speech segment, and face crop corresponds to the same emotional utterance, ready for feature extraction.

Xiangshuai Huang, “Cross-Modal Sentiment Understanding: Fusing Text, Audio, and Vision for Deeper Emotion Analysis,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 11, pp. 163-166, 2025.

B. Modality-Specific Feature Extraction

After the data undergoes preprocessing, the subsequent action is to derive features from each modality. Feature extraction indicates the procedure of converting raw input data into a collection of numerical features suitable for use in machine learning or deep learning models.

1. Text Feature Extraction: Traditional approaches for text feature extraction involve using methods like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings such as word2vec. However, more sophisticated methods, such as contextual embeddings from models like BERT, enable the extraction of more comprehensive semantic features that reflect the meaning of words in context. These embeddings portray words or sentences as dense vectors, encapsulating both syntactic and semantic details.
2. Speech Feature Extraction: In spoken communication, emotional traits can be identified through attributes such as pitch, intensity, speech rate, and MFCC. These attributes reflect a speaker's emotional state during the conversation, such as higher pitch indicating excitement or a decreased speech rate signifying sadness. Sophisticated techniques, including the use of pre-trained models like OpenSMILE or directly extracting raw features with deep neural networks (e. g., CNNs for analyzing spectrograms), have also been investigated in multimodal sentiment analysis.
3. Visual Feature Extraction: For visual data, the most commonly extracted features come from facial expressions. CNNs, particularly those pretrained on large-scale datasets such as VGG-Face, have been shown to be highly effective for facial expression recognition. The face's landmarks (eyes, mouth, eyebrows) are particularly useful for identifying emotions like happiness, surprise, anger, and sadness. Convolutional networks can likewise be employed to derive features from different visual signals, such as bodily stance, although facial expressions frequently provide the most significant information for emotion recognition.

The features obtained serve as the basis for the integration and classification phases of multimodal sentiment analysis. The quality of feature extraction is a significant factor in determining the final performance of the sentiment analysis system.

III. Fusion Strategies for Multimodal Sentiment Analysi

As multimodal sentiment analysis (MSA) systems combine different sources of information (i.e., text, speech, and visual data), one of the key challenges is how to fuse these heterogeneous modalities effectively. The integration approach is vital to the effectiveness of MSA systems, as it directly influences the model's capability to capture interactions across modalities and leverage the complementary aspects of the modalities. This section explores several advanced fusion strategies, including feature-level fusion, late fusion, hybrid fusion, and attention-based fusion, that aim to improve sentiment classification by integrating information across modalities.

A. Feature-Level Fusion

As multimodal sentiment analysis (MSA) systems integrate multiple sources of information—such as text, speech, and visual data—effectively combining these heterogeneous modalities presents a critical challenge. The choice of fusion strategy has a profound impact on the system's performance, as it directly influences the model's ability to capture cross-modal interactions and harness the complementary strengths of each modality. This section explores contemporary fusion approaches, including feature-level fusion, decision-level fusion, hybrid fusion, and attention-based fusion, each aimed at enhancing sentiment classification by merging information across modalities.

1. Concatenation: The simplest approach to feature-level fusion is concatenation, where the feature vectors from each modality (e.g., text embeddings from a transformer like BERT, speech features like MFCCs, and visual features from a CNN model) are concatenated into a single vector. This combined feature vector is then input into a classification model such as a fully connected neural network (FCNN) or an LSTM. Although this method is straightforward and effective, it can lead to very high-dimensional feature vectors, which may result in overfitting if the training data is not sufficiently large.
2. Dimensionality Reduction: To address the risk of overfitting and enhance efficiency, dimensionality reduction techniques like Principal Component Analysis (PCA) or autoencoders can be applied before concatenation. These techniques reduce the dimensionality of the feature space while retaining key information, ensuring that the fusion process does not introduce excessive redundant features.
3. Canonical Correlation Analysis (CCA): Canonical Correlation Analysis (CCA) is a statistical method that learns a shared subspace between different modalities by finding linear combinations of features that maximize their correlation. By applying CCA to the features of each modality, it is possible to learn a joint representation that effectively captures inter-modal relationships. CCA has been shown to outperform simple concatenation in certain multimodal settings by aligning the features of different modalities in a way that enhances their joint representation.

B. Decision-Level Fusion

Decision-level fusion, or late fusion, involves combining the outputs (i.e., predictions or probability distributions) of individual unimodal models after each modality has been processed independently. In this approach, each modality's model generates a sentiment prediction, which are then merged using methods like voting, weighted averaging, or stacking.

1. Voting and Averaging: A basic decision-level fusion method is majority voting, where the final sentiment

Xiangshuai Huang, "Cross-Modal Sentiment Understanding: Fusing Text, Audio, and Vision for Deeper Emotion Analysis," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 11, pp. 163-166, 2025.

prediction is based on the majority vote of the unimodal classifiers. Another approach is weighted averaging, where the outputs of the classifiers are averaged based on predefined weights, allowing more reliable modalities to contribute more to the final decision.

2. Stacking: Stacking, or stacked generalization, is a more sophisticated late fusion technique in which the predictions of unimodal models are used as inputs to a second-level classifier. This meta-classifier learns how to combine the outputs optimally based on their individual performances on the training data. Stacking allows the model to account for which modality provides more valuable information for different types of sentiment, improving overall accuracy.

Late fusion methods are advantageous in their simplicity and flexibility, as they allow for independent optimization of each modality. However, they may fail to capture the rich interactions between modalities during training, which could lead to suboptimal performance in certain scenarios.

C. Hybrid Fusion

Hybrid fusion strategies combine elements of both feature-level and decision-level fusion. These approaches aim to strike a balance by learning joint representations at intermediate stages while also preserving independent decision-making in the final layers.

1. Intermediate Fusion: In intermediate fusion, modalities are initially processed separately to extract features. At certain intermediate layers of the model, however, features from different modalities are fused. For example, after processing text features through BERT, speech features through an LSTM, and visual features through a CNN, the outputs of these intermediate layers are merged before being passed into a final decision-making module, such as a fully connected network. This method allows the model to learn both low-level interactions between features and high-level dependencies between modalities.
2. Multimodal Transformers: A powerful hybrid approach involves using multimodal transformers. Transformers, with their self-attention mechanisms, are designed to capture interactions between different parts of the input sequence. In multimodal transformers, self-attention is applied to concatenated representations of various modalities, enabling the model to learn interdependencies between text, speech, and visual data. This allows the model to capture more complex interactions, such as the relationship between words, facial expressions, and tone of voice.

Hybrid fusion strategies have been shown to provide superior performance by combining the benefits of feature-level interactions and decision-level integration of modality-specific information. However, these methods tend to be more computationally intensive, requiring additional resources for training and inference.

D. Attention-Based Fusion

Attention mechanisms, initially created for natural language processing tasks, have seen a growing application in multimodal sentiment analysis to assist the model in concentrating on the most pertinent information from each modality. Attention-based fusion methods allow the model to learn which parts of each modality are more important for sentiment classification.

1. Modality Attention: In modality attention, the model allocates varying attention weights to each modality according to its significance to the task. For instance, in a conversation, speech and facial expressions may be more indicative of sentiment than the text alone. By using a modality attention mechanism, the model can dynamically adjust the importance of each modality during the decision-making process, leading to more accurate predictions.
2. Cross-Modal Attention: Cross-modal attention mechanisms enable the model to concentrate on relevant features from one modality while taking into account features from an alternate modality. For instance, in a video, the model might pay attention to particular words in the text while also considering the related facial expression and speech tone to assess the sentiment. This cross-modal attention mechanism can be incorporated into multimodal transformers, where attention is applied not only within a single modality but also across modalities.
3. Self-Attention: Self-attention mechanisms, as used in models like the Transformer, allow the model to weigh the importance of different parts of the input sequence. In a multimodal setting, self-attention can be applied to the fused representation of text, audio, and visual data to determine which parts of the input (from any modality) should be given more focus. This allows the model to more effectively understand intricate dependencies and interactions among modalities.

IV. Experimental ResultsEditorial Policy

To validate the effectiveness of our multimodal sentiment analysis framework, we conducted experiments on several benchmark datasets widely used in the field, including CMU-MOSI, IEMOCAP, and MELD. These datasets consist of synchronized textual, acoustic, and visual data, each annotated with sentiment labels such as positive, negative, or neutral.

For feature extraction, we processed textual data using pre-trained language models for contextual embeddings, extracted acoustic features through Mel-frequency cepstral coefficients (MFCCs), and obtained visual cues using convolutional neural networks (CNNs) trained for facial emotion recognition. These modality-specific features were then jointly modeled using a multimodal transformer architecture that allows for simultaneous and contextual interaction across all modalities.

Our experimental results demonstrate that leveraging multimodal inputs leads to substantial performance gains over unimodal baselines. Specifically, the fusion of textual and acoustic modalities yielded an accuracy of 85%,

Xiangshuai Huang, “Cross-Modal Sentiment Understanding: Fusing Text, Audio, and Vision for Deeper Emotion Analysis,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 11, pp. 163-166, 2025.

outperforming text-only (75%) and speech-only (72%) configurations. The addition of visual data concerning facial expressions enhanced classification performance even more, elevating the accuracy to 88%. These outcomes underscore the effectiveness of our approach in capturing complementary sentiment signals from diverse modalities, resulting in more nuanced and reliable predictions.

V. Conclusion

In this work, we proposed a deep learning-based architecture for multimodal sentiment analysis that integrates linguistic, paralinguistic, and visual information. Our experiments highlight that combining multiple modalities significantly enhances sentiment recognition performance when compared to unimodal baselines. Particularly, the joint modeling of textual and vocal signals provides notable improvements, and incorporating facial expression data further strengthens the classifier's ability to interpret affective states.

The findings confirm the advantage of multimodal fusion in sentiment-related tasks and suggest promising directions for future research. In upcoming work, we plan to explore more sophisticated fusion mechanisms—such as hierarchical attention networks and cross-modal transformers—and test the scalability and robustness of our framework on more complex, real-world multimodal corpora.

Acknowledgment

This study was supported by the Research Innovation Fund of Anhui University of Finance and Economics (Project No. XSKY24160）

References

[1] Ramachandram D,Taylor G W.Deep multimodal learning:a survey on recent advances and trends[J].IEEE Signal Processing Magazine,2017,34(6):96-108.

[2] Habibian A,Mensink T,Snoek C G M.Video2vec embeddings recognize events when examples are scarce[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(10):2089-2103.

[3] Poria S,Cambria E,Howard N,et al.Fusing audio,visual and textual clues for sentiment analysis from multimodal content[J].Neurocomputing,2016,174:50-59.