International Journal of Information Technologies and Systems Approach
Volume 17 • Issue 1

A Multimodal Sentiment Analysis Method Integrating Multi-Layer Attention Interaction and Multi-Feature Enhancement

Shengfeng Xie, Henan Institute of Technology, China*
Jingwei Li, Henan Institute of Technology, China

ABSTRACT

To address issues related to the insufficient representation of text semantic information and the lack of deep fusion between internal modal information and intermodal information in current multimodal sentiment analysis (MSA) methods, a new method integrating multi-layer attention interaction and multi-feature enhancement (AM-MF) is proposed. First, multimodal feature extraction (MFE) is performed based on RoBERTa, ResNet, and ViT models for text, audio, and video information, and high-level features of the three modalities are obtained through self-attention mechanisms. Then, a cross modal attention (CMA) interaction module is constructed based on transformer, achieving feature fusion between different modalities. Finally, the use of a soft attention mechanism for the deep fusion of internal and intermodal information effectively achieves multimodal sentiment classification. The experimental results CH-SIMS and CMU-MOSEI datasets show that the classification results of proposed MSA method are significantly superior to other advanced comparative methods.

KEYWORDS

Cross Modal Attention, Multi-Feature Enhancement, Multi-Layer Attention Interaction, Multimodal Sentiment Analysis, Soft Attention Mechanism

INTRODUCTION

Social media provides users with convenient channels for information dissemination and collection (Pang et al., 2021; Wu et al., 2020; Yang et al., 2022; Zhang et al., 2021). With the continuous progress and development of related fields, the majority of opinions expressed on the internet now rely on various digital media technologies, including images, voice, and video, to offer more vivid and three-dimensional information content (Dayyala et al., 2022; Wen et al., 2021; Wu et al., 2022). These contents can influence the real world through dissemination and diffusion, possessing significant research value in various fields such as society, economics, politics, and others (Ahmed et al., 2022; Basiri et al., 2021; Han et al., 2021; Lai et al., 2021; Silva et al., 2022; Su et al., 2020).

Sentiment analysis refers to the process of extracting, analyzing, inductively processing, and mining subjective data with sentiment colors. One crucial task of sentiment analysis is to classify

DOI: 10.4018/IJITSA.335940 *Corresponding Author

sentiment. Early research primarily utilized single modal data for sentiment analysis, such as image, video, or text modalities (Mahabadi et al., 2021; Wan et al., 2021; Yin et al., 2022; Zhang & Yin, 2022; Zhao et al., 2021). However, when faced with massive multimodal information, although single modal data sentiment analysis has achieved success in customer satisfaction analysis and measuring voting intentions in recent years, it cannot effectively handle multimodal data due to the diversity of information, giving rise to multimodal sentiment analysis (MSA) (Cheema et al., 2021; Yang et al., 2021; Yu et al., 2021).

MSA is a computational study of viewpoints and sentiment states based on single modal sentiment analysis, using data composed of text, images, audio, or even video data. Social media is a vast source of opinions for various products and user services. The effective combination of multiple modal information can better guide analysis (Jiang et al., 2020; Li et al., 2021; Xu et al., 2022). Sentiment analysis of videos can compensate for the shortcomings of sound and vision in text sentiment analysis, and speech and facial expressions provide important clues for better recognizing the sentiment state of opinion holders. This has significant practical implications for applications such as public opinion monitoring, product recommendations, and research on user feedback (Ortis et al., 2022; Wang et al., 2020).

With the advancement of multimodal technology, contemporary academic research on sentiment analysis tasks predominantly centers around leveraging multimodal technology to enhance the accuracy of models in these tasks. However, prevailing MSA methods based on deep learning frequently encounter challenges such as inadequate representation of text semantic information, the need to balance global and local features in image modalities, and the absence of profound fusion of internal or intermodal information.

To better address the aforementioned issues and enhance the accuracy of MSA, a novel MSA method integrating multiple feature enhancements and multi-layer attention interaction is proposed. The innovation of this method, in comparison to conventional sentiment analysis approaches, can be summarized as follows:

1) For text modality, the RoBERTa model is used to extract shallow text features in the embedding layer. A representation dictionary is constructed using the Masked Language Model (MLM) to augment knowledge and enhance the semantic features of the text modality.
2) For the image modality, the ResNet and ViT models are fused to comprehensively consider both global and local image features. In addition, the incorporation of body movements, gender, and age features in video modalities enriches the feature representation of image modalities.
3) Regarding multimodal fusion, a network structure based on the deep fusion of multi-layer attention interaction is introduced. Through the multi-level interaction of the self-attention mechanism, improved Cross Modal Attention (CMA) mechanism, and soft attention mechanism, deep fusion of internal and intermodal information is achieved.

RELATED WORK

MSA has emerged as a research hotspot in recent years. Focusing on target multimodal sentiment classification and incorporating attention modules based on text and image, An et al. (2023) constructed a model capable of achieving feature fusion and extracting information correlations. On this basis, an improved universal model for target multimodal sentiment classification was proposed based on image semantic description (ITMSC). However, this method exclusively delved into an in-depth analysis of the text, neglecting fusion analysis on user expressions of different types of information. By combining an adaptive mask memory capsule network and a self-attention mechanism, sentiment analysis and clustering were constructed. Building upon this, Zhang et al. (2023) proposed the VECapsNet model and provided a corresponding MSA model based on interactive learning. However, this approach did not consider the semantic information within the image during the analysis process and lacked

detailed consideration of visual weights in multimodal fusion. Addressing the sentiment analysis problem in the Assam language, Das and Singh (2023) constructed a dedicated dataset and proposed two separate analysis modes for text and vision. By analyzing the correlation between images and text, they introduced a sentiment analysis framework for joint sentiment classification. However, this method is language-specific and has limited applicability. In response to the high error rates in automatic speech recognition affecting sentiment analysis accuracy, Wu et al. (2022) proposed a sentiment word perception multimodal refinement model (SWRM). Nevertheless, this method demonstrated inefficiency and instability, leading to a lack of robustness that requires further research. Zhang et al. (2023) conducted an analysis and calculation of multiple single mode sentiment analysis, combining them through an adaptive modal specific weight fusion network model (AdaMoW). However, this method faces limitations related to hardware constraints and network training methods, making it unsuitable for handling large-scale multimodal data. Cai et al. (2022) analyzed the novelty and applicability of four existing multimodal analysis methods for sentiment analysis, studying the suitability of different methods for various datasets. They summarized and compared the development trends and model performance but did not propose a novel MSA method on this basis. In response to the absence of MSA methods for a specific small language, Das and Singh (2023) proposed an MSA framework based on hybrid fusion. They constructed a single analysis method by conducting separate semantic and visual analyses on images and text. A corresponding joint sentiment analysis model was then constructed by fusing the two forms of information. However, this method solely considers both image and text information, falling short in effectively analyzing the increasing volume of available video information.

Based on this analysis, existing MSA methods grounded in deep learning often encounter issues such as insufficient representation of text semantic information, difficulty in balancing global and local features of image modalities, and a lack of internal modal information or deep fusion between modalities. To address these challenges, the proposed method utilizes feature enhancement to rectify the insufficient representation of text semantic information. By combining ResNet and ViT to balance the global and local features of image modalities, a multi-layer attention interaction structure is designed to achieve deep fusion of information within or between modalities.

PROPOSED MSA MODEL

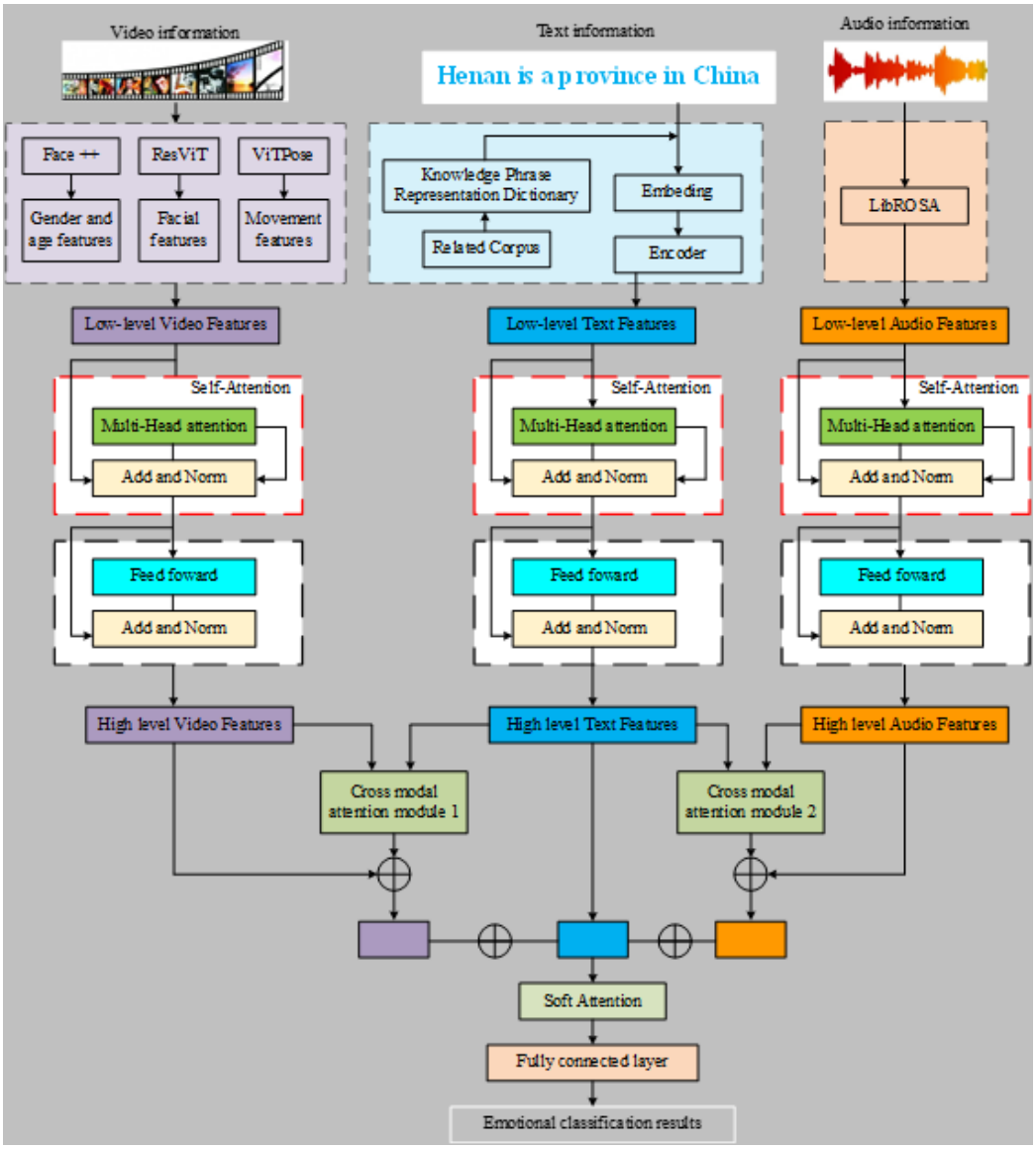
After investigating and thoroughly analyzing current MSA methods, a new MSA model based on attention mechanisms and multiple features (AM-MF) is proposed. The AM-MF primarily comprises three parts: Multimodal Feature Extraction (MFE), multi-layer attention fusion, and MSA. The overall architecture diagram of the AM-MF is depicted in Figure 1.

In Figure 1, the AM-MF acquires rich low-level features from each modality through an MFE module. It utilizes a self-attention mechanism to extract internal information from the three modalities, generating corresponding high-level features. Subsequently, interaction is facilitated through CMA mechanisms, enabling the spatial exchange of information between modalities. The acquired internal and intermodal information is concatenated to obtain more comprehensive audio and video fusion features. Finally, the final representations of modalities are fed into a soft attention module, which assigns varying weights to the three modalities, ultimately achieving multimodal sentiment classification through a fully connected layer. An in-depth introduction to the modeling process of the AM-MF will be presented in the sequel.

MFE

In the MFE module, the features of audio, text, and video modalities are extracted through their respective sub-networks and converted into vector representations comprehensible to deep neural networks for learning. Various extraction methods fall into the following categories:

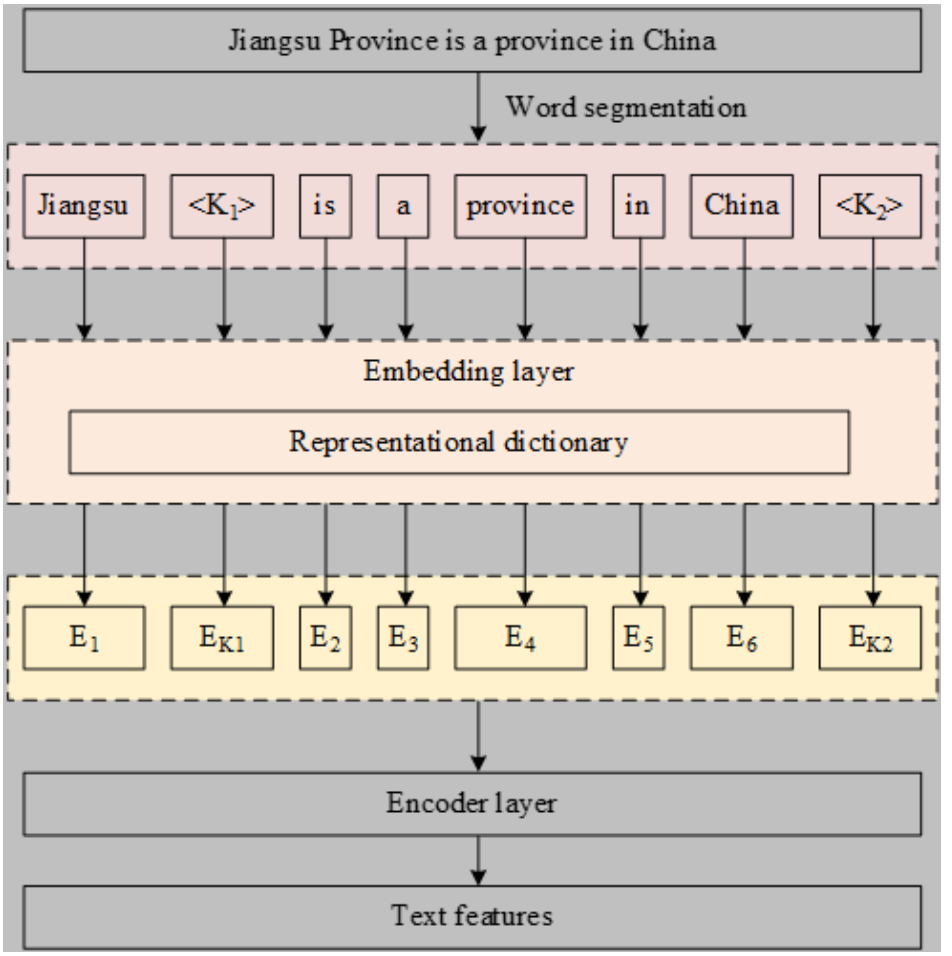
Figure 1. Structure of the AM-MF



1) For the text modality, the RoBERTa model is used to extract text features, and knowledge enhancement is conducted through a representation dictionary. The Embedding layer transforms words into embeddings, which are then input into the Encoder layer for encoding to obtain text features. Knowledge enhancement is carried out in the Embedding layer. The RoBERTa model is illustrated in Figure 2.

In Figure 2, the text underwent initial preprocessing to remove meaningless special symbols, URLs, and other unnecessary representations. Subsequently, word segmentation is applied to the text. The segmented text is then transformed into word embeddings through the Embedding layer.

Figure 2. Structure of the text feature extraction module



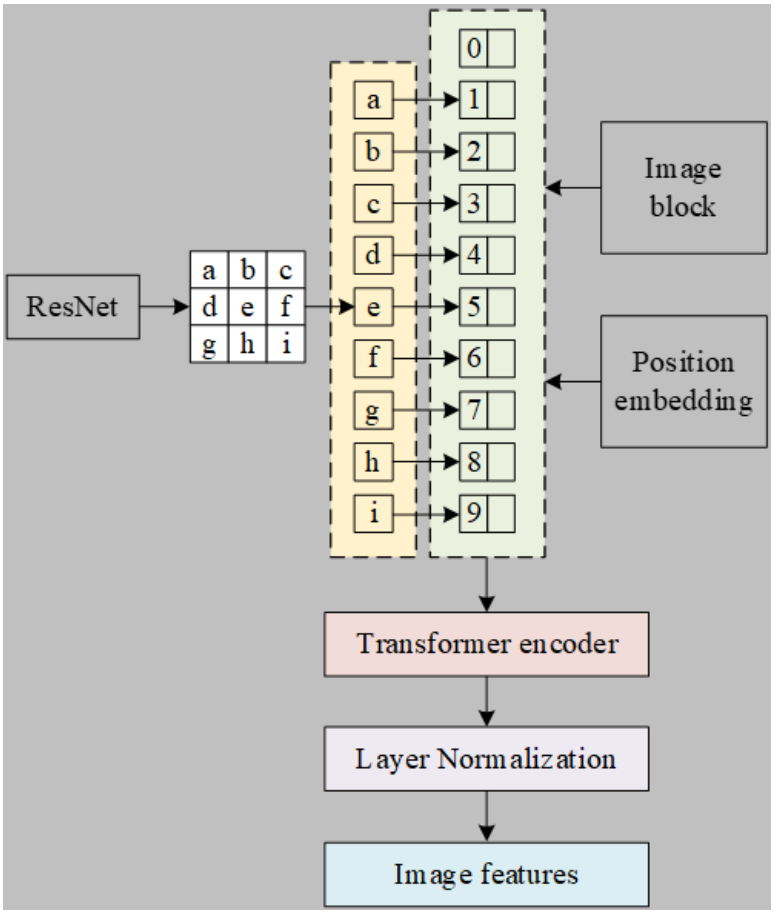
The original word embeddings of sentences are learned through the mask language task during the pre-training of the RoBERTa model.

2) For the video modality, ResViT is employed to extract high-level features of the image modality. ResViT is illustrated in Figure 3.

In Figure 3, the structure of ResViT includes two parts: ResNet and ViT based on ImageNet pre-training. Using the ResViT toolkit, facial symbols, facial action units, head direction, gaze direction, and other information are extracted to obtain the facial features of the video.

In addition, ViTPose, a state-of-the-art scheme in human pose estimation, leverages the efficient learning ability of ViT to obtain high-quality features. In this context, the pre-trained ViTPose on the MS COCO key point dataset is used to extract the action features of characters in the video. Initially, all video clips undergo frame extraction processing. Extracted frames are fixed and evenly distributed, automatically selecting frame extraction intervals based on the video duration to create a set. Passing all extracted frames into ViTPose yields five key point coordinates of the head, upper limbs, and lower limbs. The resulting vector dynamically reflects the information of the character's action transformation in the video, serving as the character's action feature. In addition, gender and

Figure 3. Structure of ResViT



age characteristics are extracted using the Open Vision Face++artificial intelligence open platform. Finally, the facial features, motion features, gender, and age features of the video are input into a linear transformation layer to achieve uniform dimensions and are concatenated as low-level video modal features.

3) For the audio modality, the LibROSA speech toolkit is used to extract acoustic features of 22050HZ, thereby obtaining low-level audio modality features.

Multilayer Attention Fusion

Effectively fusing multimodal features has consistently been the primary challenge in MSA. The advent of attention mechanisms has prompted researchers to enhance MSA models by incorporating these mechanisms. Transformers, when applied to single mode data, excel in capturing global contextual relationships, continually reinforcing the internal structure of the single mode, and obtaining more robust self-feature information. Increasingly, researchers are leveraging Transformers to accomplish cross-modal feature fusion on multimodal data. However, Transformers exhibit certain drawbacks. The model possesses a complex structure and a substantial number of parameters, and the fusion of multimodal features demands extensive training time. Moreover, Transformer-based fusion models

often overlook important information within a single modality, leading to insufficient information fusion between modalities. Studies indicate that sentiment classification performance is superior in text modality compared to audio and video. In the cross-modal interaction module, the text modality aids in modeling audio and video modalities, achieving cross-modal fusion through an improved CMA mechanism.

Self-Attention

To begin, the Transformer's capability to capture contextual relationships is harnessed to model single-mode low-level features, aiming to acquire more intricate high-level feature information. For text, the feature representation F_k^L of the k-th video is fed into the Transformer, employing multi-head self-attention to learn the modality's internal information, as expressed in Eq. (1) - (4):

$$\begin{cases} Q_L = F_k^L \omega_Q \\ K_L = F_k^L \omega_K \\ V_L = F_k^L \omega_V \end{cases}, \tag{1}$$

$$Att\left(Q_L, K_L, V_L\right) = softmax\left(\frac{Q_L K_L^D}{\sqrt{\lambda_k}}\right) V_L \tag{2}$$

where $\omega_Q \in R^{\lambda_k^L \times \lambda_k}$, $\omega_K \in R^{\lambda_k^L \times \lambda_k}$, and $\omega_V \in R^{\lambda_k^L \times \lambda_v}$ are the linear transformation weight matrices of text features, and $\lambda_k^L = \lambda_k = \lambda_v$ is the corresponding dimension size,

$$H_h = Att\left(Q_L \omega_h^Q, K_L \omega_h^K, V_L \omega_h^V\right) \tag{3}$$

$$MH\left(Q_L, K_L, V_L\right) = Co\left(H_1, H_2, ..., H_h\right) \omega_h^Z \tag{4}$$

where $\omega_h^Q \in R^{\lambda_k^L \times \lambda_k^h}$, $\omega_h^K \in R^{\lambda_k^L \times \lambda_k^h}$, $\omega_h^V \in R^{\lambda_k^L \times \lambda_v^h}$, and $\omega_h^Z \in R^{\lambda_k^L \times h\lambda_v}$ are the linear transformation weight matrices of text features in multi-head attention, respectively, and $\lambda_k^h = \lambda_v^h = \frac{\lambda_k^L}{h}$ is the dimension size of each head.

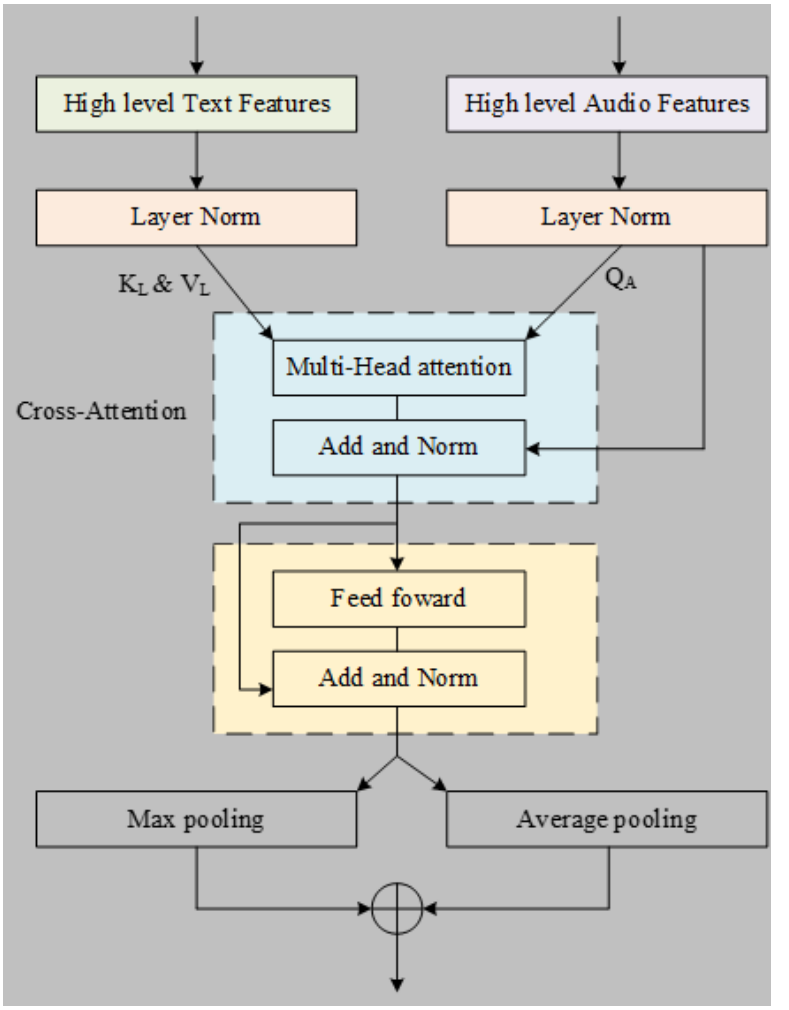
Following the multi-head self-attention mechanism, a vector representation of the internal relationships within the text modality is obtained through residual connection and layer normalization. Subsequently, a feedforward neural network consisting of two linear layers is applied, ultimately yielding a high-level text feature representation F_k^L. Similarly, high-level audio feature representation F_k^A and high-level video feature representation F_k^V can be derived.

CMA

In the CMA, cross-modal fusion is accomplished by enhancing the Transformer. The improved Transformer can accept two modes as inputs, integrating the high-level text feature representation F_k^L and high-level audio feature representation F_k^A into the CMA module. Here, F_k^A is the main mode providing Q, and F_k^L serves as an auxiliary mode, providing K and V. The CMA structure is illustrated in Figure 4.

The CMA using text-assisted audio is represented as:

Figure 4. Structure of CMA



$$\begin{aligned} \underset{L\to A}{CrossModal} &= softmax\left(\frac{Q_L K_L^D}{\sqrt{\lambda_k}}\right)V_L \\ &= softmax\left(\frac{F_k^A \omega_{QA} F_k^{L^D} \omega_{KL}^D}{\sqrt{\lambda_k}}\right)F_k^A \omega_{VL} \end{aligned} \tag{5}$$

where $\omega_{QA} \in R^{\lambda_k^A \times \lambda_k}$, $\omega_{KL} \in R^{\lambda_k^L \times \lambda_k}$, and $\omega_{VL} \in R^{\lambda_k^L \times \lambda_v}$ are the linear transformation weight matrices.

Following cross-modal multi-head attention, feature vectors that fuse text and audio modal information are obtained through residual connections and layer normalization operations, achieving interactive fusion of information between modalities. After passing through a feedforward neural network composed of two linear layers, the audio feature vector fused with text feature information

$F_k^{L\to A}$ is ultimately obtained through residual connection and layer normalization. The use of pooling operations is advantageous for suppressing noise, reducing information redundancy, computational efficiency, and preventing overfitting. Maximum pooling captures local features at each moment, while average pooling concentrates the model on global features. Combining both pooling types results in richer feature layers. Splicing the outcomes of maximum and average pooling together yields the CMA's output:

$$\begin{cases} F_{k_{\max}}^{L\to A} = maxpooling\left(F_k^{L\to A}\right) \\ F_{k_{avg}}^{L\to A} = averagepooling\left(F_k^{L\to A}\right) \\ F_k^{L\to A} = Concat\left(F_{k_{\max}}^{L\to A}, F_{k_{avg}}^{L\to A}\right) \end{cases} \tag{6}$$

To achieve the fusion of internal and interaction information between single modalities, the high-level audio and video features within the modalities are concatenated with the corresponding features of cross-modal fusion:

$$\begin{cases} U_k^A = Concat\left(F_k^A, F_k^{L\to A}\right) \\ U_k^V = Concat\left(F_k^V, F_k^{L\to V}\right) \end{cases} \tag{7}$$

Using a linear transformation layer, the audio and video features are dimensionally reduced to match the text features. Subsequently, the three modal features are concatenated to form the final multimodal feature representation:

$$U_k = Concat\left(F_k^L, U_k^A, U_k^V\right) \tag{8}$$

Modal Fusion Using Soft Attention

In various contexts, each modality does not equally contribute the final sentiment expression. From the perspective of textual modality alone, a video discourse may express only a neutral sentiment. However, considering the speed, intonation of the audio, and facial expressions of the characters in the video, it may convey a negative or positive sentiment. In MSA, it is essential to weigh the contribution of different modalities. Therefore, a soft attention mechanism module is introduced before completing modal information fusion for classification, assigning varying weights to different modalities. Due to the distinct contributions of each modality to the task, the AM-MF uses modal fusion in this module rather than directly connecting features of different modalities.

The eigenvector of each mode $m\left(m \in \left\{text, video, audio\right\}\right)$ is recorded as Z_m and converted into a fixed length form Z'_m, and a two-layer feedforward neural network is then used to calculate the attention weights of each mode. A weighted average is employed to obtain a single, fixed-length fusion feature Z_{fu}. The calculation process is expressed as follows:

$$\begin{cases} \omega'_m = \omega_{m2} \cdot \tanh\left(\omega_{m1} \cdot Z_m + b_{m1}\right) + b_{m2} \\ \omega_m = softamx\left(\omega'_m\right) \\ Z'_m = \tanh\left(\omega_{m3} \cdot Z_m + b_{m3}\right) \\ Z_{fu} = \sum_{m \in \{text, image\}} \omega_m \cdot Z'_m \end{cases} \tag{9}$$

where ω_{m1}, ω_{m2}, and ω_{m3} are the weights that can be obtained from model training, b_{m1}, b_{m2}, and b_{m3} are the bias values that can be obtained from model training, ω_m is the self-attention weight, and Z_{fu} is the final feature representation.

Classification Training

Using the SoftMax classifier for sentiment classification:

$$c = softamx\left(\omega_{multi} \cdot Z_{fu} + b_{multi}\right) \tag{10}$$

The model is trained using the minimum cross entropy loss function:

$$L = -\sum_{k=1}^{N} c'_k \log\left(c_k\right) \tag{11}$$

where c is the distribution of sentiment labels and N is the category of emotions.

EXPERIMENT AND ANALYSIS

Experimental Environment

The experiment was conducted on the server side of the system with Ubuntu version 18.04 and specific hardware configurations outlined in Table 1.

Experimental Dataset

Two different datasets were used during the experiment, as follows:

1) CH-SIMS dataset (Yu et al., 2020). This dataset comprises Chinese single modal and MSA data, including 2,281 video clips. For multimodal datasets, a typical feature is that characters in

Table 1. Hardware configuration of experimental platform

Parameters	Configuration
OS	Linux
CPU	Intel(R) Xeon(R) Gold 5118 CPU
CPU Memory	16G @ 2.30GHz
GPU	Tesla V100
Programming Language	Python 3.8.13
Programming environment	PyTorch 1.12.1
CUDA	11.4

the video must emit sound while also having facial segments to obtain corresponding features. To meet these requirements, 3,210 fragments were collected from over 100 source videos for experimentation.

2) CMU-MOSEI dataset (Zadeh et al., 2019). This dataset is one of the largest three-mode datasets, containing 23,453 video clips and 250 sound clips from 1,000 characters. Each video clip is accompanied by a corresponding explanation. In addition, the dataset includes two labels: sentiment and emotion. There are a total of seven categories of sentiments from negative to positive, and the values of the labels are between -3 and 3. The dataset provides raw data, and for text, audio, and video files, their images need to be captured spontaneously at a fixed frequency.

The aforementioned datasets have been divided in a ratio of 6:2:2, with 60% used for training, while the validation and testing sets each account for 20%. Each dataset is categorized into negative, neutral, and positive sentiments, as detailed in Table 2.

The meanings of symbols in Table 2 are as follows: NG-Negative, WN-Weak Negative, UN-Neutral, WP-Weak Positive, PS-Positive.

Evaluation Indicators and Parameter Settings

In the experiment, the following three evaluation indicators were used to assess the performance of different models:

1) Acc x. Acc-x represents the accuracy of class x classification, with a larger value indicating better performance. Specifically, Acc-2 represents NG and PS, Acc-3 represents NG, UN, and PS, Acc-5 represents NG, WN, UN, WP, and PS, and Acc-7 represents strong negative, NG, WN, UN, WP, PS, and strongly positive sentiment information.
2) Weighted average F1 Score. F1 Score considers both accuracy and recall comprehensively, with a larger value indicating better performance.
3) Mean Absolute Error (MAE). MAE is the average of the absolute error values between the calculated true and predicted values, with lower values indicating better performance.

Table 2. Statistical information of the dataset

		Training Set	Verification Set	Test Set
CH-SIMS	Total	1369	456	456
	NG	449	150	150
	WN	300	100	100
	UN	200	70	70
	WP	120	70	70
	PS	300	66	66
CMU-MOSEI	Total	14071	4691	4691
	NG	4761	1591	1591
	WN	3000	1200	1200
	UN	2110	700	700
	WP	1700	400	400
	PS	2500	800	800

The experiment was conducted on the server side. The key to the model's good performance lies in setting reasonable model parameters, as shown in Table 3.

Hyperparameter Analysis

Loss Training

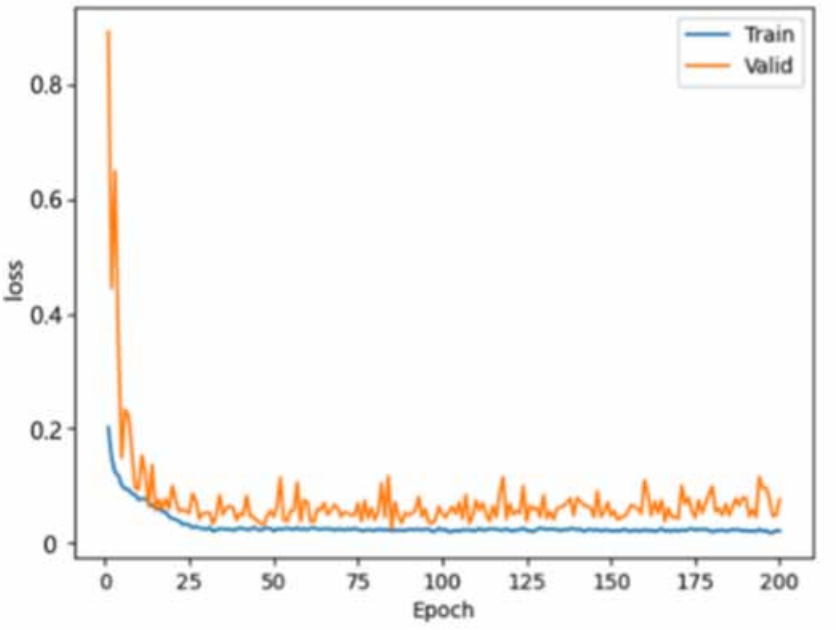
The CH-SIMS dataset is used to train the proposed AM-MF, and the model losses during the training process are observed and recorded. The results are shown in Figure 5.

In Figure 5, after 30 training epochs, the loss values of the proposed AM-MF are minimized, and the model has converged on both the training and validation sets.

Table 3. Parameter settings for the model

Parameters	Value
CMA dimension	50
Number of CMA heads	10
Optimizer	Adam
Epochs	30
Learning rate	0.001
Dropout	0.4
Early stop	8
Batch size	16

Figure 5. Training loss of AM-MF



Learning Rate Training

Taking the CH-SIMS dataset as an example, an experimental analysis was conducted on the performance of the proposed AM-MF under different learning rates. The experimental results of the AM-MF under different learning rates are shown in Figure 6.

In Figure 6, the results of Acc-2, Acc-3, Acc-5, and F1 Score show a continuous upward trend during the learning rate change from 0.1 to 0.001. However, as the learning rate changes from 0.001 to 0.00001, the values of the above evaluation indicators begin to show a downward trend. The proposed model achieves optimal performance when the learning rate is set to 0.001. Hence, in the subsequent experimental phase, the learning rate for the proposed models will be set to 0.001.

Dropout Analysis

During the model training process, the setting of Dropout can affect the training results. An experimental analysis was conducted on the performance of the proposed AM-MF under different Dropouts using the CH-SIMS dataset. The experimental results are presented in Figure 7.

In Figure 7, during the process of Dropout changing from 0.1 to 0.4, the outcomes of Acc-2, Acc-3, Acc-5, and F1-Score show a continuous upward trend. However, during the process of Dropout changing from 0.4 to 0.8, these evaluation indicators begin to show a downward trend. The proposed model can achieve optimal performance when Dropout is set to 0.4. Therefore, in the subsequent experimental phase, the Dropout for the proposed models will be set to 0.4.

Result Comparison and Analysis

To further verify the superiority of the proposed AM-MF, a comparative analysis was conducted on CH-SIMS and CMU-MOSEI using the AM-MF, ITMSC [24], VECapsNet [25], and SWRM [27] models.

Figure 6. Learning rate influence on the AM-MF

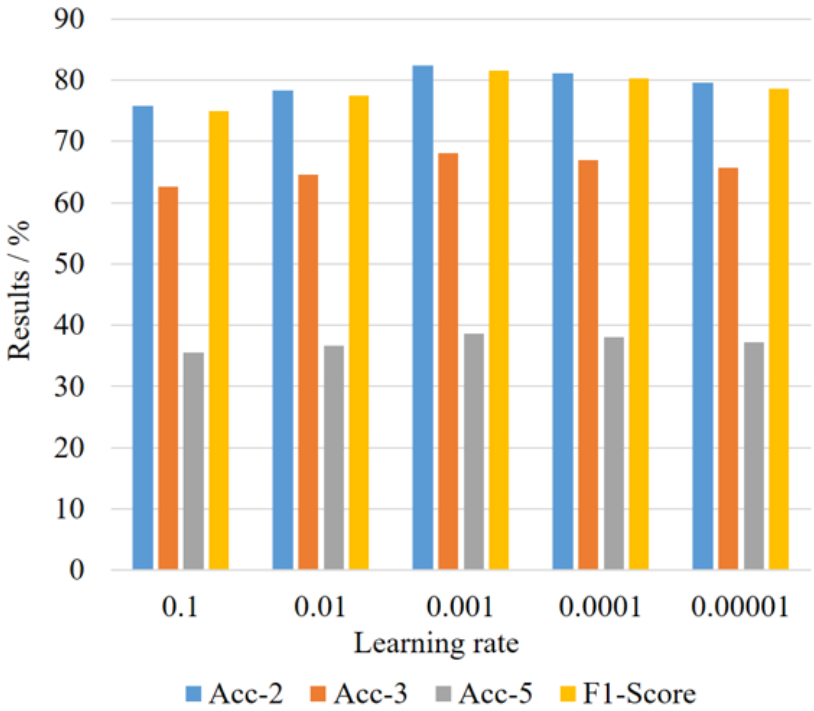
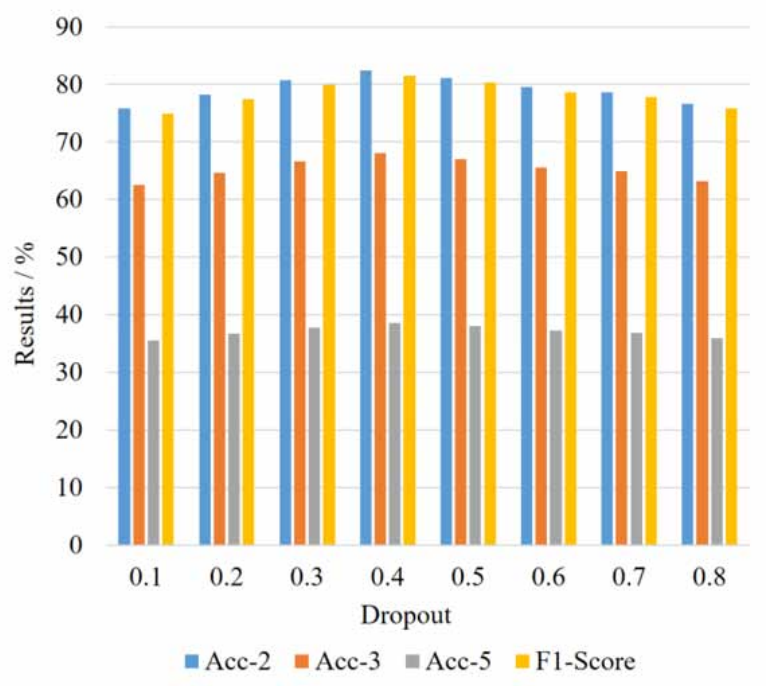


Figure 7. Dropout influence on the AM-MF



Under the same experimental conditions, different evaluation index values calculated by different models using the CH-SIMS dataset are shown in Figure 8 and Table 4.

Similarly, under the same experimental conditions, different evaluation index values calculated by different models when using the CMU-MOSEI dataset are shown in Figure 9 and Table 5.

The MAE values obtained through various methods in the experiment with different datasets are illustrated in Figure 10.

In the experimental results, the Acc-x and F1 Score of the proposed AM-MF are both the largest, and MAE values are the smallest when using different datasets. With the CH-SIMS dataset, Acc-2, Acc-3, Acc-5, and F1 Score of the AM-MF were 82.39%, 68.02%, 38.62%, and 81.50%, respectively, with an MAE value of 0.402. When using the CMU-MOSEI dataset, these performance indicator values become 76.68%, 44.92%, 44.76%, and 76.83%, respectively, with an MAE value of 0.72. This is attributed to the utilization of the RoBERTa model, which extracts shallow text features in the embedding layer, significantly augmenting the semantic aspects of the text modality. In addition, the model, relying on the multi-level interaction of the self-attention mechanism, improved CMA mechanism, and soft attention mechanism, achieves deep information fusion within and between modalities, thereby substantially enhancing the accuracy of sentiment classification.

Ablation Experiment

Modal Ablation

Using the CH-SIMS dataset as an example, different modal experiments were conducted with T, A, and V representing the text, audio, and video modalities. The results are shown in Table 6.

In Table 6, the proposed AM-MF model proves effective for MSA tasks under various modal combinations. Notably, combining all three models yielded the best results. Moreover, when lacking the text modality, the proposed AM-MF model exhibited the poorest performance, underscoring

Figure 8. Comparison of different models in CH-SIMS

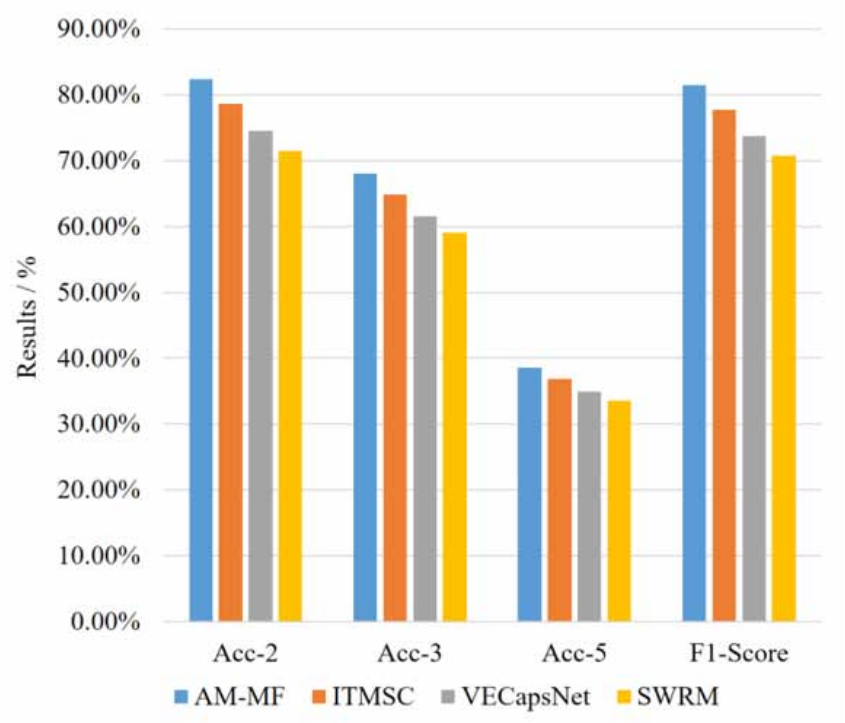


Table 4. Comparison of different models in CH-SIMS

Model Indicator	AM-MF	ITMSC	VECapsNet	SWRM
Acc-2	82.39%	78.60%	74.56%	71.51%
Acc-3	68.02%	64.89%	61.56%	59.04%
Acc-5	38.62%	36.84%	34.95%	33.52%
F1-Score	81.50%	77.75%	73.76%	70.74%
MAE	0.402	0.471	0.502	0.624

the significant impact of the text modality on enhancing sentiment analysis methods, as it can more intuitively represent emotional information.

Model Ablation

To further illustrate the importance and distinctions of each part of the model, ablation experiments were conducted on the proposed AM-MF using the CH-SIMS dataset. Simulation analysis included AM, MF, AM-MF without fusion and complete AM-MFs. Additionally, the number of floating-point operations per second (FLOPs) for several models and the time required for predicting a single sample were provided to reflect the real-time performance of each model. Definitions: AM represents an MSA model with only attention mechanisms, MF represents an MSA model with only MFE, and AM-MF without fusion represents an AM-MF model lacking information fusion. The obtained results are presented in Table 7.

Figure 9. Comparison of different models in the CMU-MOSEI

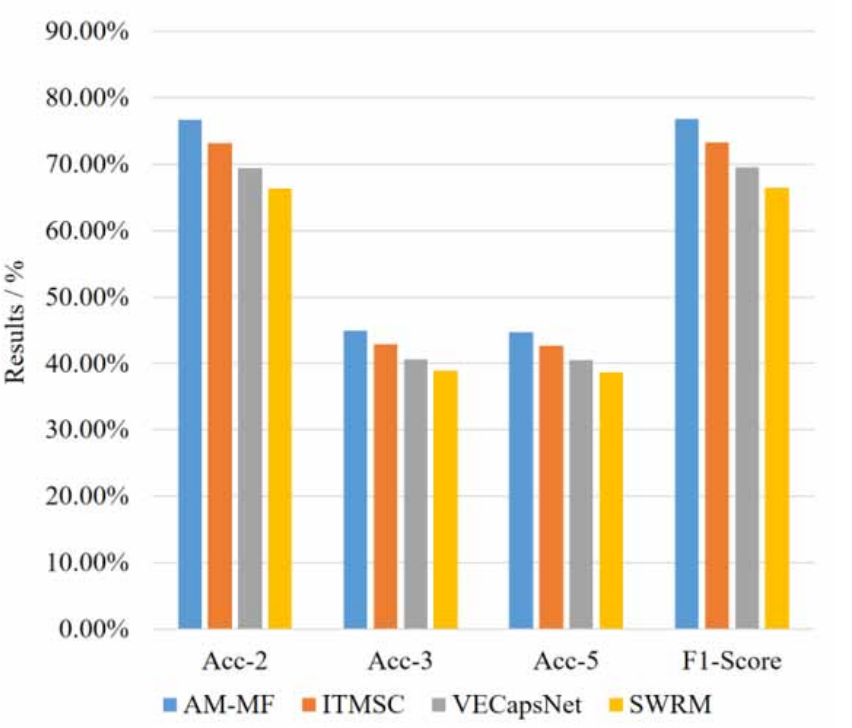


Table 5 Comparison of different models in the CMU-MOSEI

Model Indicator	AM-MF	ITMSC	VECapsNet	SWRM
Acc-2	76.68%	73.15%	69.40%	66.33%
Acc-5	44.92%	42.85%	40.65%	38.86%
Acc-7	44.76%	42.70%	40.51%	38.72%
F1-Score	76.83%	73.30%	69.53%	66.46%
MAE	0.72	0.83	0.96	1.10

The ablation experimental results in Table 7 indicate that the AM model slightly outperforms the MF model, suggesting a slightly greater impact of the attention mechanism than MFE. Moreover, AM-MF without fusion outperforms AM and MF models, emphasizing the significant roles of attention mechanisms and MFE compared to information fusion. However, the introduction of information fusion further enhances the performance of the proposed model. Despite the AM-MF model exhibiting slightly inferior performance in terms of FLOPs and prediction time compared to other ablation models, its integration of AM and MF, which increases its complexity to some extent, positions it as the optimal choice based on accuracy and prediction time.

CONCLUSION

The proposed MSA method, integrating multi-feature enhancement and multi-layer attention interaction, is proposed to address key issues prevalent in current MSA approaches, such as

Figure 10. MAE values for different models with different datasets

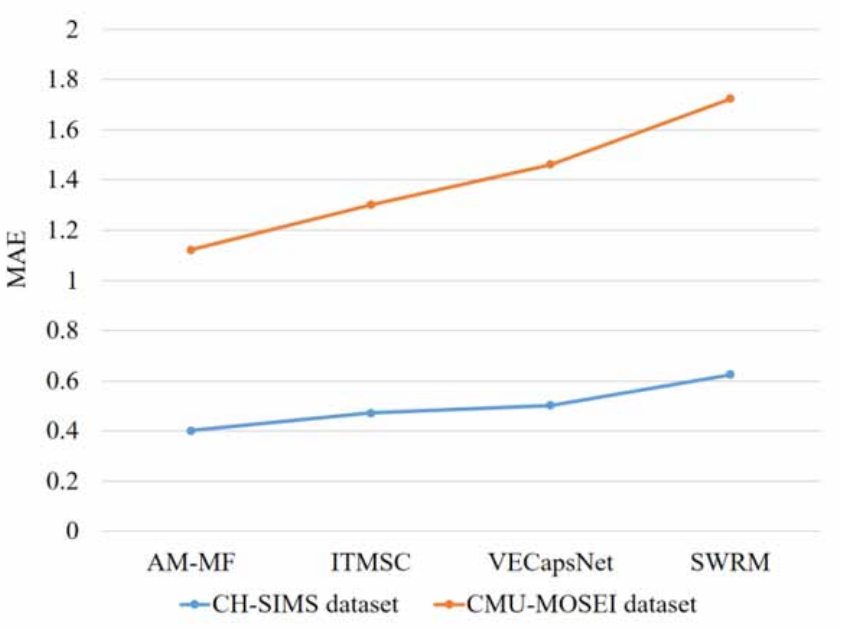


Table 6. Modal ablation experiment results of AM-MF

Modals	Acc-2	Acc-3	Acc-5	F1-Score	MAE
T, A	80.82%	67.15%	37.68%	80.38%	0.439
T, V	80.39%	67.60%	37.42%	80.77%	0.436
A, V	79.83%	65.25%	36.08%	78.69%	0.492
T, A, V	82.39%	68.02%	38.62%	81.50%	0.402

Table 7. Model ablation experiment results of AM-MF

Model	Acc-2	Acc-3	Acc-5	F1-Score	MAE	FLOPs /10^6	Prediction Time/s
AM	77.96%	63.15%	35.04%	78.29%	0.558	2.518	0.023
MF	79.44%	64.59%	36.45%	78.83%	0.576	2.605	0.023
AM-MF without fusion	80.09%	66.51%	37.79%	80.31%	0.480	2.726	0.029
AM-MF	82.39%	68.02%	38.62%	81.50%	0.402	2.893	0.032

inadequate representation of text semantic information, challenges in balancing global and local features of image modalities, and the absence of deep fusion of information within or between modalities. Experimental results underscore the effectiveness of the proposed method in tackling these challenges. The introduction of the RoBERTa model into the embedding layer during the extraction of text information significantly enhances the semantic features of the text modality. Furthermore, the incorporation of ResNet and ViT during the extraction of modal features from data containing audio and video information allows for a comprehensive consideration of global and local features,

thereby enriching the feature representation of modalities. The network structure, based on deep fusion of multi-layer attention interactions through multi-level interaction involving the self-attention mechanism, improved CMA mechanism, and soft attention mechanism, contributes substantially to the enhancement of multimodal sentiment classification accuracy.

However, the deep fusion of features within and between modalities, relying on the Transformer, has somewhat increased the number of model parameters and manual parameter adjustment workload. Future advancements will involve the incorporation of heuristic algorithms for adaptive neural search, enabling automatic adjustments of the model's structure and parameters based on changes in the environment and data, thereby adapting to diverse tasks and data distributions. To further improve the model, various attention mechanisms and enhancement techniques will be introduced. Additionally, exploration into sentiment factors and subjects hidden in multimodal data will be conducted to augment the model's interpretability in terms of sentiment understanding.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are included within the article.

CONFLICTS OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this paper.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the authors of the article.

REFERENCES

Ahmed, F., Faraz, N. A., Ahmad, N., & Iqbal, M. K. (2022). Supportive leadership and post-adoption use of MOOCs: The mediating role of innovative work behavior. *Journal of Organizational and End User Computing*, *34*(1), 1–23. doi:10.4018/JOEUC.308813

An, J., Zainon, N. W., Mohd, W., & Hao, Z. (2023). Improving targeted multimodal sentiment classification with semantic description of images. *Computers, Materials & Continua*, *75*(3), 5801–5815. doi:10.32604/cmc.2023.038220

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharrya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems*, *115*(2), 279–294. doi:10.1016/j.future.2020.08.005

Cai, Z., Gao, H., Li, J., & Wang, X. (2022). Deep learning approaches on multimodal sentiment analysis. *IEEE International Conference on Electrical Engineering, Big Data and Algorithms, (EEBDA).* doi:10.1109/EEBDA53927.2022.9745018

Cheema, G. S., Hakimov, S., Müller-Budack, E., & Ewerth, R. (2021). A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding,* 37-45. doi:10.1145/3463945.3469058

Das, R., & Singh, T. D. (2023). A hybrid fusion-based machine learning framework to improve sentiment prediction of assamese in low resource setting. *Multimedia Tools and Applications*, *12*(3), 74–82. doi:10.1007/s11042-023-15356-3

Das, R., & Singh, T. D. (2023). Image-text multimodal sentiment analysis framework of Assamese news articles using late fusion. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(6), 382–390. doi:10.1145/3584861

Dayyala, N., Walstrom, K. A., Bagchi, K. K., & Udo, G. (2022). Factors impacting defect density in software development projects. *International Journal of Information Technologies and Systems Approach*, *15*(1), 1–23. doi:10.4018/IJITSA.304813

Han, W., Chen, H., & Poria, S. (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. ACL*, 9180–9192. doi:10.18653/v1/2021.emnlp-main.723

Jiang, T., Wang, J., Liu, Z., & Ling, Y. (2020). Fusion-extraction network for multimodal sentiment analysis. *Lecture Notes in Computer Science*, *12*(5), 785–793. doi:10.1007/978-3-030-47436-2_59

Lai, X., Wu, X., & Zhang, L. (2021). Autoencoder-based multi-task learning for imputation and classification of incomplete data. *Applied Soft Computing*, *98*, 1–13. doi:10.1016/j.asoc.2020.106838

Li, S., Liu, Z., & Li, Y. (2021). Temporal and spatial evolution of online public sentiment on emergencies. *Information Processing & Management*, *57*(2), 102177–102191. doi:10.1016/j.ipm.2019.102177 PMID:32287939

Mahabadi, R. K., Belinkov, Y., & Henderson, J. (2021). Variational information bottleneck for effective low-resource fine-tuning. *Proceedings of the 9th International Conference on Learning Representations. ICLR*, 1–13.

Ortis, A., Farinella, G. M., & Battiato, S. (2022). Survey on visual sentiment analysis. *IET Image Processing*, *14*(8), 1440–1456. doi:10.1049/iet-ipr.2019.1270

Pang, J., Rao, Y., Xie, H., Wang, X., Wang, F. L., Wong, T.-L., & Li, Q. (2021). Fast supervised topic models for short text emotion detection. *IEEE Transactions on Cybernetics*, *51*(2), 815–828. doi:10.1109/TCYB.2019.2940520 PMID:31567111

Silva, A. P., & Marques, R. P. (2022). The contribution of ERP systems to the maturity of internal audits. *International Journal of Information Technologies and Systems Approach*, *15*(1), 1–25. doi:10.4018/IJITSA.311501

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. V. B. (2020, April 26-30). *VL-BERT: Pre-training of generic visual-linguistic representations* [Conference presentation]. 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia.

Wan, Z., Zhang, C., Zhu, P., & Hu, Q. (2021). Multi-view information-bottleneck representation learning. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 10085–10092.

Wang, J., Yu, L.-C., Lai, K. R., & Zhang, X. (2020). Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*(4), 581–591. doi:10.1109/TASLP.2019.2959251

Wen, S., Wei, H., Yang, Y., Guo, Z., Zeng, Z., Huang, T., & Chen, Y. (2021). Memristive LSTM network for sentiment analysis. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, *51*(3), 1794–1804.

Wu, L., Qi, M., Jian, M., & Zhang, H. (2020). Visual sentiment analysis by combining global and local information. *Neural Processing Letters*, *51*(3), 2063–2075. doi:10.1007/s11063-019-10027-7

Wu, Y., Zhao, Y., Yang, H., & Chen, S. (2022). Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. *60th Annual Meeting of the Association-for-Computational-Linguistics (ACL),* 1397-1406.

Xu, F., Pan, Z., & Xia, R. (2022). E-commerce product review sentiment classification based on a naive Bayes continuous learning framework. *Information Processing & Management*, *57*(5), 102221–102228. doi:10.1016/j.ipm.2020.102221

Yang, X., Feng, S., Wang, D., & Zhang, Y. (2021). Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, *23*(5), 4014–4026. doi:10.1109/TMM.2020.3035277

Yang, Y., Siau, K., Xie, W., & Sun, Y. (2022). Smart health: Intelligent healthcare systems in the metaverse, artificial intelligence, and data science era. *Journal of Organizational and End User Computing*, *34*(1), 1–14. doi:10.4018/JOEUC.308814

Yin, C., Zhang, S., Wang, J., & Xiong, N. N. (2022). Anomaly detection based on convolutional recurrent autoencoder for IoT time series. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, *52*(1), 112–122. doi:10.1109/TSMC.2020.2968516

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., & Yang, K. (2020). CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 343-348. doi:10.18653/v1/2020.acl-main.343

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 10790–10797. doi:10.1609/aaai.v35i12.17289

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *56th Annual Meeting of the Association-for-Computational-Linguistics (ACL)*, 2236-2246.

Zhang, J. L., Wu, X. M., & Huang, C. Q. (2023). AdaMoW: Multimodal sentiment analysis based on adaptive modality-specific weight fusion network. *IEEE Access : Practical Innovations, Open Solutions*, *11*(4), 48410–48420. doi:10.1109/ACCESS.2023.3276932

Zhang, S., Li, B., & Yin, C. (2022). Cross-modal sentiment sensing with visual-augmented representation and diverse decision fusion. *Sensors (Basel)*, *22*(1), 74–82. doi:10.3390/s22010074 PMID:35009620

Zhang, Y., Tiwari, P., Song, D., Mao, X., Wang, P., Li, X., & Pandey, H. M. (2021). Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis. *Neural Networks*, *133*, 40–56. doi:10.1016/j.neunet.2020.10.001 PMID:33125917

Zhang, Y. F., Zhang, Z. Q., Feng, S., & Wang, D. (2023). Visual enhancement capsule network for aspect-based multimodal sentiment analysis. *Applied Sciences-Basel*, *12*(23), 239–250.

Zhao, L., Wang, Y., & Zhao, J. (2021). Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12793–12802. doi:10.1109/CVPR46437.2021.01260

Shengfeng Xie, Lecturer, master. He graduated from Henan Polytechnic University in 2010. Worked in Henan Institute of Technology His research interests include Deep Learning and The Internet of Things.

Jingwei Li, associate professor, master. He graduated from Harbin Institute of Technology in 2005. Worked in Henan Institute of Technology. His research interests include big data and cloud computing and computer network.