International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1

Multimodal Sentiment Analysis Method Based on Hierarchical Adaptive Feature Fusion Network

Huchao Zhang, Zhejiang Agricultural Business College, Shaoxing, China*

ABSTRACT

The traditional multi-modal sentiment analysis (MSA) method usually considers the multi-modal characteristics to be equally important and ignores the contribution of different modes to the final MSA result. Therefore, an MSA method based on hierarchical adaptive feature fusion network is proposed. Firstly, RoBERTa, ResViT, and LibROSA are used to extract different modal features and construct a layered adaptive multi-modal fusion network. Then, the multi-modal feature extraction module and cross-modal feature interaction module are combined to realize the interactive fusion of information between modes. Finally, an adaptive gating mechanism is introduced to design a global multi-modal feature interaction module to learn the unique features of different modes. The experimental results on three public data sets show that the proposed method can make full use of multi-modal information, outperform other advanced comparison methods, improve the accuracy and robustness of sentiment analysis, and is expected to achieve better results in the field of sentiment analysis.

KEYWORDS

Adaptive Gating Mechanism, Cross-Modal Feature Interaction, Hierarchy Adaptation, Multimodal Feature Fusion, Sentiment Analysis

INTRODUCTION

Social media, as a network platform for users to create, share, and communicate, can make it more convenient for users to access information and also provide them with more choices and editing rights. Unlike paper media, such as newspapers, social media has a variety of content forms (Sahoo & Gupta, 2021; Ahmed et al., 2022; Almomani et al., 2022). In addition to text mode, social media can also provide users with more intuitive and three-dimensional information content through modes such as voice and image. Images, speech, and text constitute the most common scenes in daily life (Su et al., 2023; Gao et al., 2022; Balcilar et al., 2021).

Sentiments play a crucial role in our daily lives, helping us communicate, learn, and make decisions. For a long time, researchers have been dedicated to using machines to analyze human sentiments (Tiwari et al., 2021; Schneider et al., 2023; Singh & Sachan, 2021). Early MSA often focused on single modality information such as sound, text, visual, and biological signals. However,

DOI: 10.4018/IJSWIS.335918 *Corresponding Author

using a single modality for MSA users often did not accurately analyze their sentiments (Salhi et al., 2021; Mohammed et al., 2022; Garcia-Garcia, 2023). Because the same text may express opposite meanings in different contexts, it is difficult to accurately predict users' sentiments based solely on one modality. Due to the inability of single-mode MSA technology to effectively process data and fully utilize the diversity of information, it is no longer suitable for the current complex environment (Sun et al., 2020; Yuan et al., 2021; Zhang et al., 2022).

As research deepens, researchers have found that information can more effectively analyze human sentiments than single-modal information. The way people communicate and express sentiments in daily life is usually through the fusion of sound, text, and visual modalities. MSA is research on mining user perspectives and emotional states from data, such as text, vision, or speech, based on single-mode MSA (Chen et al., 2022; Niu et al., 2021; Poria et al., 2023).

Multimodal emotion recognition can be used to analyze user emotions on social media. By combining text, image, and video data, users' emotional tendencies and emotional states can be more accurately understood.In the field of education, multimodal plays an important role. Multimodal emotion recognition analyzes students' speech, facial expression, and gesture data and can understand students' emotional state and learning effect and provide personalized teaching and feedback. When integrating a multi-modal emotion recognition system, it can be integrated with an existing system or platform through an API interface or SDK. The results of multimodal emotion recognition can be used as input for decision-making, personalized recommendation, emotion analysis, and other functions.

Analyzing the sentiment information of users in multimodal data can better help cloud service providers and enterprises grasp the emotional state of users and guide their next steps (Kumar & Sivakumar, 2022; Guo et al., 2022). For example, users' comments on e-commerce products, to a certain extent, reflect their level of preference for the product, and these comments will have a significant impact on potential users of the product. If enterprises can promptly improve their products in response to negative comments, they can avoid significant economic losses. MSA can better guide the analysis of sentiment classifiers (Yadav & Vishwakarma, 2023). MSA of videos can compensate for the shortcomings of sound and visuals in text SA. Voice and facial expressions provide important clues for better identifying the emotional state of opinion holders, which has significant practical uses for research on user feedback (Zhang et al., 2022; Liao et al., 2022).

The MSA algorithm based on has more advantages in robustness and accuracy, and it has gradually become the mainstream of MSA research. In addition, the cross-modal hierarchical will better model dynamics between and within modalities, which is of great significance for studying machine learning. However, existing fusion methods often have some significant issues, such as 1) Insufficient single-mode high-level feature extraction. 2) Treating multimodal features as equally important and focusing more on the direct fusion of multimodal features, while neglecting the contribution of different modalities to the final MSA, leading to insufficient utilization of important modal information. 3) It is difficult to balance local modal-related features and global modal unique features, resulting in the loss of important features and affecting the performance of MSA.

A level adaptive fusion method based on text modality guidance is proposed to address these issues. Compared with traditional MSA methods, the innovation of this method lies in:

1) In the phase of modal feature extraction, the advantage of the in capturing contextual relationships is utilized to model single modal low-level features and obtain richer high-level feature information.
2) To address the issue of insufficient information fusion between modalities, a local cross-modal interaction module was designed. Using the text modality with a greater degree of contribution as the guiding modality, and the speech and visual modalities with a lesser degree of contribution as the auxiliary modality, the cross-modal attention is utilized to achieve the representation of important information.

3) To better learn the unique features of different modalities, a global multimodal feature interaction module was designed, and hierarchical adaptive fusion based on important information was achieved through the adaptive gating mechanism.
4) A local-global feature fusion module was designed, which combines local modal related features and global modal unique features to achieve comprehensive judgment of sentiment.

RELATED WORKS

Today, people are willing to publish user-generated videos to express their sentiments and viewpoints. This provides a large number of data sources for MSA. MSA aims to use multimodal information to analyze human sentiment, which has become the main focus of sentiment computing research. Many MSA methods have been proposed and achieved excellent results, which have excellent performance and robustness compared to single-mode SA. At present, the existing MSA methods can be divided into nontext mode-dominated MSA and text mode-dominated MSA, according to the different dominant modes. The following will introduce them separately.

Non Modal Dominated MSA

In the non-textual modal-dominated MSA method, a multi-label training scheme was designed in reference (Yu et al., 2021), which can generate additional single modal labels for each modality and train simultaneously with the main task. The deep neural network method avoids complex feature engineering and can adaptively learn data features. However, the internal structure of deep neural networks is relatively complex and has poor interpretability.

By using a cross-modal converter to map other modes to the target mode, reference (Zhang et al., 2022) proposes an integrated consistency difference network, on which multiple single modal labels are obtained through self-supervision for MSA tasks. However, this method cannot effectively fuse multiple single-mode labels and obtain analysis results based on this. Mai et al., (2020) propose an adversarial codec framework that achieves multimodal MSA by transforming the feature distribution of the source mode into the feature distribution of the target mode. However, the proposed framework structure is relatively complex, resulting in high computational complexity and low MSA efficiency. Yang et al., (2022) propose a translation framework that improves the quality of BERT. However, this method can only extract text information from other modalities, and it cannot comprehensively extract features of tone in audio and expressions and actions in video.

Text Modal Dominated MSA

In the MSA method dominated by text modality, Wang et al., (2020) proposes an end-to-end translation network, which uses transform between modalities and captures the correlation between multimodal features through forward and backward translation to improve translation performance. However, this method can only achieve good results when applied to text information, and it is less applicable to other types of fusion information. Considering the poor quality of non-natural language emotional features, it will weaken emotions during feature fusion. Wang et al., (2022) propose a modal reinforcement cross-attention module.

Modal translation methods can improve modal quality, but this method requires processing multiple forms of input, requiring more complex models and higher computational resources. By calculating the bimodal attention matrix between two different modes, Huddar et al., (2020) spliced it into a three-mode attention matrix to fuse the interaction information between different modes, and proposed an effective MSA method. However, the fusion mechanism of this method has shortcomings in cross-modal modeling and cannot capture the connections between multiple modalities well. Xi et al., (2020) utilize self-attention and multi-head interactive attention to get correlations between different modalities, which can improve the accuracy. However, after the introduction of multi-level

attention mechanisms, this method needed to calculate the attention weights of each input position, which increased the computational complexity when processing large-scale data.

It can be seen from the previous work that non-text modality-dominated multimodal emotion recognition mainly focuses on non-text data, such as images, audio, video, etc. This recognition also uses specially designed feature extraction methods, such as image feature extraction, audio feature extraction, etc. Multimodal emotion recognition mainly focuses on text data. When extracting features, it mainly focuses on extracting text features, such as word embedding and text vectorization. The hierarchical adaptive feature fusion network emotion recognition is a comprehensive multi-modal data method, which can process both text and non-text data. Text and non-text features are extracted and fused at different levels to capture semantic and visual information of the data more comprehensively.

Based on the above analysis, existing fusion methods usually focus on the direct fusion of multimodal features, while neglecting the contribution of different modalities to emotional features, which can easily cause important information loss and affect the performance of MSA. To this end, a level adaptive fusion method based on text modality guidance is proposed, which can fully consider the feature interaction between modalities, while also taking into account local modality-related features and global modality-specific features, effectively improving sentiment classification performance.

PROPOSED MULTIMODAL

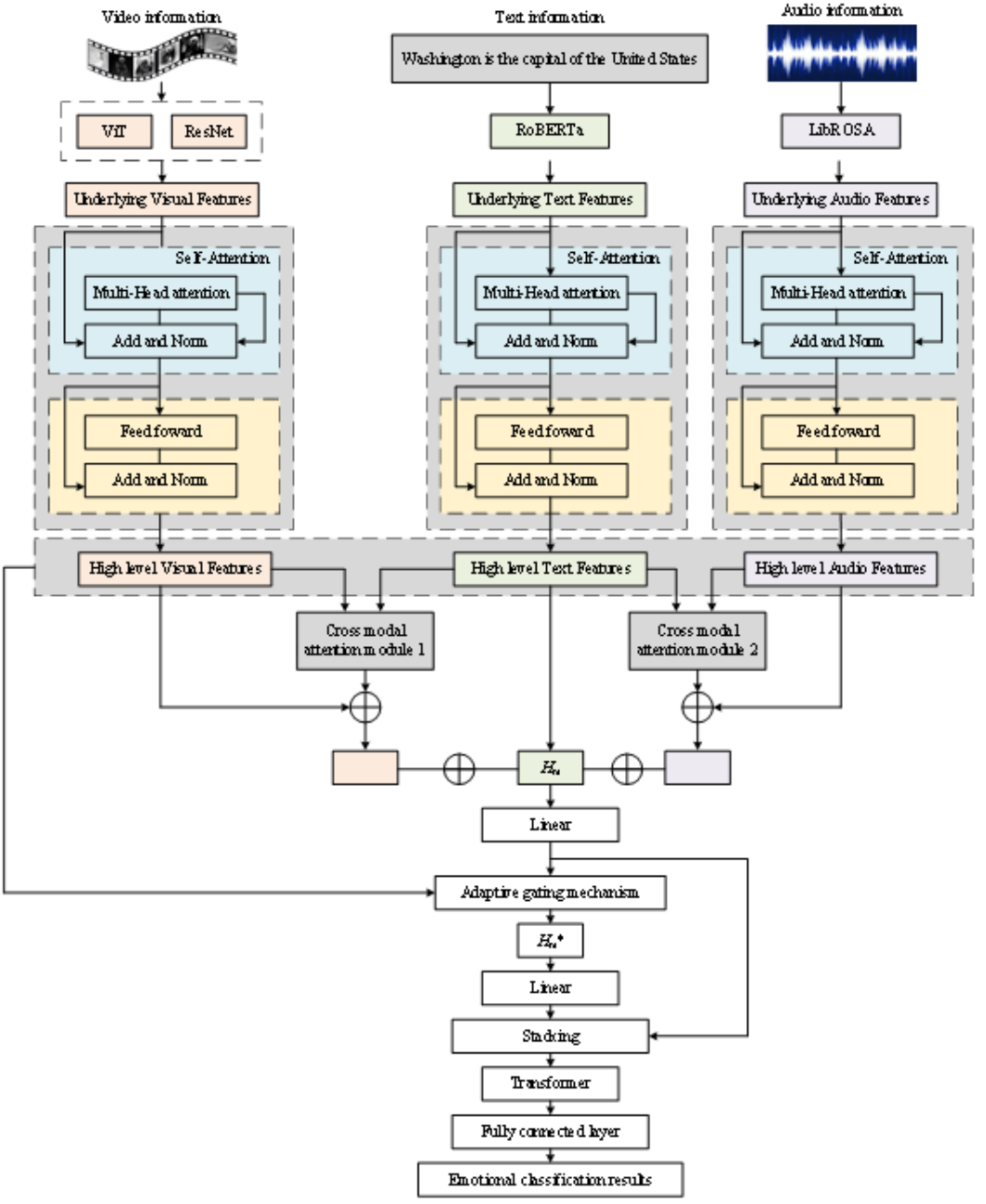
For text modality, the RoBERTa is used, and knowledge enhancement is performed through a representation dictionary. The rationale for using RoBERTa for feature extraction is to pre-train the deep bidirectional model using large-scale unsupervised data to obtain a diverse and universal representation of the language. In RoBERTa, the Embedding layer converts words into words for embedding and then inputs them into the Encoder layer for encoding to obtain text features. Knowledge enhancement is carried out in the Embedding layer.

For video modality, ResViT is used to extract high-level semantic features of image modality. ResViT (Vision Transformer) is a new vision pre-training model that uses a encoder to extract image features. Unlike traditional convolutional neural networks (CNNS), ResViT uses multi-head self-attention mechanisms to model the relationships between different locations in an image to obtain a more global representation of visual features. The structure of ResViT includes two parts: ResNet and ViT based on ImageNet pre-training. By using the ResViT toolkit to extract facial symbols, facial action units, head direction, gaze direction, and other information, the facial features of the video are obtained.

For audio modality, use the LibROSA speech toolkit to extract the acoustic of 22050HZ, then obtain low-level audio modality features. LibROSA can extract time domain features (such as time domain envelope, short-time energy, etc.), spectral features (such as Meir spectrum cepstrum coefficient (MFCC), Meir frequency spectra, etc.), and other advanced features (such as tone, rhythm, etc.) of audio signals.

On this basis, a hierarchical adaptive fusion network model was constructed. The overall architecture diagram is shown in Fig. 1. There is a close relationship between each component of the model. Firstly, the multimodal feature extraction module is the foundation of the whole system, which is responsible for extracting feature representation from input data of different modes. Next, the output of the multimodal feature extraction module is passed to the multimodal feature extraction module. The goal of this module is to further learn and extract multimodal features to obtain more representative feature representations and model and fuse multimodal features. At the same time, the adaptive gating mechanism can adjust the feature weights of different modes adaptively to better integrate the global multi-modal information. Finally, the local-global feature fusion module is used to realize the feature interaction and fusion between different modes. In summary, each module works with each other to achieve effective processing and performance improvement of multi-modal emotion recognition tasks.

Figure 1. Adaptive multimodal



In Figure 1,linear is a linear unit that is a fundamental component of deep learning. “Stacking” refers to stacking multiple neural network layers in turn to form a deeper neural network model.

Multimodal

For video mode, assuming there is a total of *M* videos, each containing *m* discourse segments, the *k*-th video can be represented as $V_k = V_{k1}, V_{k2}, V_{k3}, ..., V_{km}$. Pass the text, audio, and video of the *l*-th

segment of the *k*-th video into their respective single mode feature extraction modules to obtain corresponding text feature $F_{kl}^{\ L}$, speech feature $F_{kl}^{\ A}$, and video feature $F_{kl}^{\ V}$. The features of the *k*-th video are represented as follows:

$$F_k^n = \left[f_{k1}^{\ n}, f_{k2}^{\ n}, f_{k3}^{\ n}, ..., f_{km}^{\ n}\right] \tag{1}$$

In eq(1), $F_k^n \in R^{H_k \times D_k^n}$, $n \in \left[K, V, L\right]$,*L* is text mode, *A* is audio mode, and *V* is video mode. H_k is discourse amount contained in the *k*-th video, D_k^n is the feature dimensions of each modality in the *k*-th video.

Firstly, the advantage of in capturing contextual relationships is utilized to model single-mode low-level features and obtain richer high-level feature information. Taking the text modal features as an example, input the text feature F_k^L of the *k*-th video into:

$$\begin{cases} Q_L = F_k^L \omega_Q \\ K_L = F_k^L \omega_K \\ V_L = F_k^L \omega_V \end{cases} \tag{2}$$

$$Att\left(Q_L, K_L, V_L\right) = softmax\left(\frac{Q_L K_L^D}{\sqrt{\lambda_k}}\right) V_L \tag{3}$$

In eq(2) and eq(3), $\omega_Q \in R^{\lambda_k^L \times \lambda_k}$, $\omega_K \in R^{\lambda_k^L \times \lambda_k}$, $\omega_V \in R^{\lambda_k^L \times \lambda_v}$ are linear transformation weight matrices for text features, respectively. $\lambda_k^L = \lambda_k = \lambda_v$ Q is the corresponding dimension size.

$$H_h = Att\left(Q_L \omega_h^Q, K_L \omega_h^K, V_L \omega_h^V\right) \tag{4}$$

$$MH\left(Q_L, K_L, V_L\right) = Co\left(H_1, H_2, ..., H_h\right) \omega_h^Z \tag{5}$$

In eq(4) and eq(5), $\omega_h^Q \in R^{\lambda_k^L \times \lambda_k^h}$, $\omega_h^K \in R^{\lambda_k^L \times \lambda_k^h}$, $\omega_h^V \in R^{\lambda_k^L \times \lambda_v^h}$, $\omega_h^Z \in R^{\lambda_k^L \times h\lambda_v}$ is the linear transformation weight matrices of text features in multi-head attention. $\lambda_k^h = \lambda_v^h = \frac{\lambda_k^L}{h}$ is the dimensional size of each head.

Then, a vector representation of internal relationships of the text modality is obtained through residual connection and layer normalization operations. Then, a feedforward neural network composed of two linear layers is formed, and finally, a high-level text feature Tab. $H_L \in R^{l_L \times \lambda_L}$ is obtained through residual connection and layer normalization. l_L represent the sequence length, and λ_L represent the feature dimension. Similarly, high-level audio feature representation $H_A \in R^{l_A \times \lambda_A}$ and high-level video feature representation $H_V \in R^{l_V \times \lambda_V}$ can be obtained, where l_A and l_V represent the sequence lengths of audio and video modalities, respectively, and λ_A and λ_V represent the feature dimensions of video modalities.

Cross-Modal Feature Interaction

In cross-modal feature interaction, cross-modal fusion is achieved through an improved . The improved can receive two modalities as inputs, and it inputs the high-level text feature representation H_L and

high-level audio feature representation H_A together into the cross-modal feature interaction module. As the main mode, H_A provides Q. H_L serves as an auxiliary mode, providing K and V. The structure of it is seen in Figure 2.

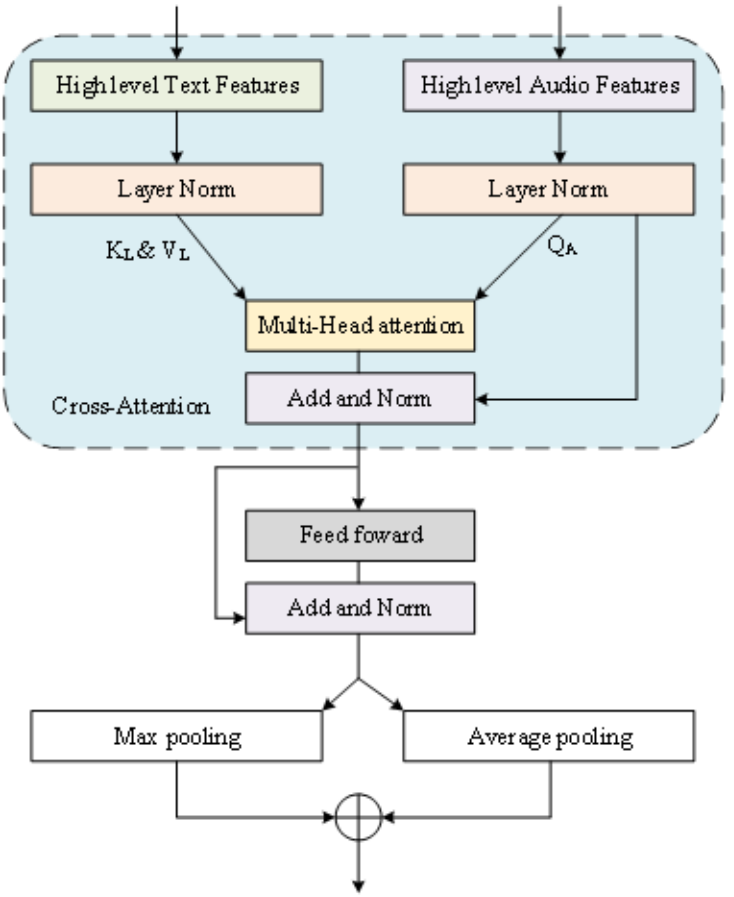
The cross-modal feature interaction representation using text-assisted audio is as follows:

$$\begin{aligned} \underset{L\to A}{CrossModal} &= softmax\left(\frac{Q_L K_L^D}{\sqrt{\lambda_k}}\right)V_L \\ &= softmax\left(\frac{H_k^A \omega_{QA} H_k^{L^D} \omega_{KL}^D}{\sqrt{\lambda_k}}\right)H_k^A \omega_{VL} \end{aligned} \tag{6}$$

In eq(6), D is the words amount in the k-th audio, $\omega_{QA} \in R^{\lambda_k^A \times \lambda_k}$, $\omega_{KL} \in R^{\lambda_k^L \times \lambda_k}$, $\omega_{VL} \in R^{\lambda_k^L \times \lambda_v}$ linear transformation weights。

After cross-modal multi-head attention, the feature vectors that fuse text and audio modal information are obtained through residual connections and layer normalization operations, enabling

Figure 2. Structure of module



full learning of the information of two modalities and achieving interactive fusion of information between modalities. After passing through a feedforward neural network composed of two linear layers, and finally, through residual connection and layer normalization, the audio feature vector $F_k^{L\to A}$ fused with text feature information is obtained.

Due to the advantages of pooling operation in suppressing noise, reducing information redundancy, model computation, and preventing overfitting, combined pooling is chosen to obtain richer feature layers, maximum pooling is chosen to capture local features at each moment, and average pooling is chosen to make the model more focused on global features. Splice the results of maximum pooling and average pooling together as the output:

$$\begin{cases} F_{k_{\max}}^{L\to A} = maxpooling\left(F_k^{L\to A}\right) \\ F_{k_{avg}}^{L\to A} = averagepooling\left(F_k^{L\to A}\right) \\ F_k^{L\to A} = Concat\left(F_{k_{\max}}^{L\to A}, F_{k_{avg}}^{L\to A}\right) \end{cases} \tag{7}$$

To get the fusion of internal information 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:

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

Through a linear transformation layer, the audio and video features are dimensionally reduced to match the text feature dimensions, and then the three modal features are concatenated together as the final multimodal feature representation:

$$U_k = Concat\left(F_k^L, U_k^A, U_k^V\right) \tag{9}$$

Then connect text features H_L, text voice interaction features H_A, and text visual interaction features H_V, and map them to a low dimensional space, as follows:

$$H_m = \mathrm{ReLU}\left\{\omega_{l1}^{mL}\left[H_L; H_L^A; H_L^V\right] + B_{l1}^m\right\} \tag{10}$$

In eq(10), $\omega_{l1}^{mL} \in R^{\left(\lambda_L+\lambda_A+\lambda_V\right)\times\lambda_m}$, ReLU is the activation number, and H_m represents the local correlation features of the three modes.

Adaptive

A global multimodal feature interaction module was designed using an adaptive gating unit. In the feature integration part, the features of each mode are weighted and averaged by using the weight vector obtained by the adaptive gating mechanism. This preserves important information about each mode feature and weakens the impact of irrelevant or noisy information. The weighted average features are input to the next layer for subsequent processing and task learning. This module is guided by relevant

features mainly based on text modality, and it uses a gating mechanism to obtain unique features of three modalities, taking speech modality as an example. Its network structure is seen in Figure 3.

In the global multimodal feature interaction module, the output modal-related features H_m of the local modal interaction module and the output speech modal features H_A of the feature representation module are inputted into two independent linear layers, respectively. The outputs of the two linear layers are used as inputs to the gating unit, and the unique features of a single modality are filtered out using the features. The process is as follows:

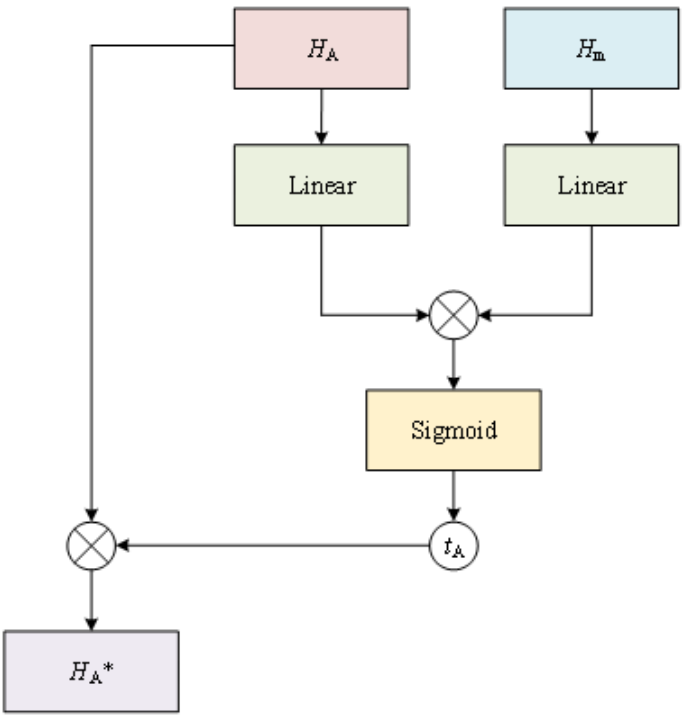
$$\begin{cases} t_A = Sigmoid\left(\omega_m H_m + \omega_A H_A\right) \\ H_A{}^* = \left(1 - t_A\right) * H_A \end{cases} \tag{11}$$

In the formula, t_A is the similarity weight between related features and speech features, ω_m and ω_A are parameter matrices, and $H_A{}^* \in R^{l_A \times \lambda_A}$ is the unique feature of speech modality. Repeating the above steps can obtain unique features of text modality and visual modality, represented as $H_L{}^* \in R^{l_L \times \lambda_L}$ and $H_V{}^* \in R^{l_V \times \lambda_V}$.

Then connect text-specific features $H_L{}^*$, speech-specific features $H_A{}^*$, and visual-specific features $H_V{}^*$, and map them to low dimensional space R^{λ_m}. The process is as follows:

$$H_m{}^* = \mathrm{ReLU}\left\{\omega_{l2}{}^{mL}\left[H_L{}^*; H_A{}^*; H_V{}^*\right] + B_{l1}{}^m\right\} \tag{12}$$

Figure 3. Structure of feature interaction



In eq(12), $\omega_{l2}^{mL} \in R^{\left(\lambda_L+\lambda_A+\lambda_V\right)\times\lambda_m}$, ReLU is the activation number, H_m^* which is the globally unique features of the three modes.

MSA

Through the local cross-modal interaction module, local modal-related features H_m guided by text modality are obtained. Through the global interaction module, global modal unique features H_m^* are obtained using relevant features mainly guided by text modality. To integrate multimodal features and modal importance information to better achieve MSA, a local-global feature fusion module was designed . Firstly, the modal-related features and modal-specific features are added to the matrix $G=\left[H_m, H_m^*\right] \in R^{2\times\lambda_m}$. Then, the matrix R is used as the input, each vector learns other cross-modal representations, integrating global and local features to achieve a comprehensive judgment of sentiment.

For the self-attention, define $Q=K=V=G \in R^{2\times\lambda_m}$, and the generates a new matrix $G'=\left[H_m', H'^*_m\right]$, with the calculation process as follows:

$$Atten\left(Q,K,V\right)=Softmax\left(\frac{Q_L K_L^D}{\sqrt{\lambda_k}}\right)V_L \tag{13}$$

$$Head_k = Atten\left(Q\omega_i^q, K\omega_i^k, V\omega_i^v\right) \tag{14}$$

$$\begin{aligned} G' &= MultiHead\left(G;\beta^{atten}\right) \\ &= \omega^o\left(Head_1 \oplus Head_2 \oplus ... \oplus Head_n\right) \end{aligned} \tag{15}$$

In eq(13), (14), and (15), $\omega_i^{q,k,v} \in R^{\lambda_m\times\lambda_m}$, ω^o is linear transformation weight, $\oplus$ represents stitching, $\beta^{atten}=\left[\omega^q,\omega^k,\omega^v,\omega^o\right]$

Finally, send output vectors into the linear layer to obtain the result:

$$H_o=\left[H_m';H'^*_m\right] \tag{16}$$

$$Y=\omega_{l3}^m H_o + B_{l3}^m \tag{17}$$

In eq(16) and (17), Y is the final prediction result, with a sentiment score of [-3,+3], H_m' is a modal related feature, H'^*_m is a modal specific feature, $H_o \in R^{2\times\lambda_m}$, $\omega_{l3}^m \in R^{\lambda_m\times1}$, λ_m is a low dimensional spatial dimension.

Using binary cross entropy loss as the loss function, the model is optimized by minimizing the cross entropy between the predicted output in the training sample and the actual sample true value, achieving the final MSA task. The function is represented as:

$$Loss=-\sum\nolimits_n\left[Y^{(i)}\log\left(Y^{(i)}\right)+\left(1-Y^{(i)}\right)\log\left(1-Y^{(i)}\right)\right] \tag{18}$$

EXPERIMENT

Experimental

This experiment was run on the server side of the system version Ubuntu 18.04 as shown in Table 1.

Table 1. Experimental

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

Experimental

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

(1) CH-SIMS dataset (Xi et al., 2020). This dataset is a Chinese single modal and MSA dataset. For the multimodal dataset, the typical feature is that the characters in the video must have facial segments to obtain corresponding features while making sounds. Finally, 3210 pieces of data were collected.
(2) CMU-MOSEI dataset (Zadeh et al., 2017). It includes 23453 video data and 250 audio data. In addition, the dataset has two labels: sentiments and emotions, with a total of 7 categories, and the values of the labels range from [-3 to 3]. The dataset provides raw data, and for text, audio, and video files, their images need to be captured spontaneously at a fixed frequency.
(3) MELD dataset (Kamath et al., 2007) is a dataset for multimodal sentiment analysis developed at Stanford University. MELD dataset contains video, audio, and text data from movie conversations. The conversations in the dataset covered different emotional categories, including joy, sadness, anger, fear, disgust, and neutral. Each dialogue contains interactions between multiple characters, as well as their facial expressions, voice, and text messages. The experimental datasets are shown in Table 2.

Experimental

In the experiment, the following three evaluation indicators were used to test models:

(1)Accuracy (Acc).

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{19}$$

Table 2. Experimental

Dataset	CH-SIMS	CMU-MOSEI	MELD
Training Set	1284	16300	8050
Verification Set	230	1900	1000
Test Set	685	4653	1980
Total number	2199	22856	10407

(2)F1-Score.

$$F1-Score=\frac{2TP}{2TP+FN+FP} \tag{20}$$

The meanings of the symbols in the equation are depicted in Table 3.

(3)Mean Absolute Error (MAE).

$$MAE=\frac{1}{M}\sum_{k=1}^{M}\left|Y_k-Y_{k0}\right| \tag{21}$$

(4)Correlation coefficient between predicted value and true value (Corr)

$$Corr=\mathrm{cov}\left(Y_k,Y_{k0}\right)\cdot\sigma Y_k\cdot\sigma Y_{k0} \tag{22}$$

In the equation, Y_k and Y_{k0} are the predicted and true values, respectively. $\mathrm{cov}\left(Y_k,Y_{k0}\right)$ are the covariance of Y_k and Y_{k0}. σY_k and σY_{k0} are the standard deviations of Y_k and Y_{k0}, respectively.

The parameters of the model during the experiment are depicted in Table 4.

During modal fusion, the weights between the modes are adjusted according to the quality and reliability of the data. For more noisy or incomplete modes, reduce their weight to reduce their shadow on the final result. Moreover, the models are stacked by ensemble learning, which reduces the impact of noise and incomplete data by integrating the predictions of multiple models and improves the robustness and accuracy of the models.

Table 3. Confusion matrix

		Prediction	
		1	**0**
Actual	1	T_P	F_N
	0	F_P	T_N

Table 4. Parameter

Parameters	CH-SIMS	CMU-MOSEI	MELD
Optimizer	Adam	Adam	Adam
Activation	ReLU	ReLU	ReLU
Learning rate	0.002	0.001	0.001
Learning rate (BERT)	5×10-5	5×10-5	5×10-5
Batch size	32	32	32
Dropout	0.4	0.3	0.4
Epoch	30	25	25
λm	128	128	128

Comparative

Comparative analysis will be conducted under the same experimental conditions based on the CH-SIMS and CMU-MOSEI for the proposed model and various other different models.

Comparative With CMU-MOSEI

Firstly, based on the CMU-MOSEI dataset, a comparative analysis was conducted between the proposed model, the MAG-BERT model (Rahman et al., 2020), and the UniMSE model (Hu et al., 2022). Under the same experimental conditions, different evaluation index values calculated by different models are shown in Figure 4 and Table 5.

In Figure 4 and Table 5, when using the CMU-MOSEI dataset, the proposed model has the largest Acc, F1 Score, and Corr, reaching 0.8627, 0.8610, and 0.7760, respectively, while the MAE is the smallest, at 0.5210. This means that the proposed model has better performance compared to MAG-BERT and UniMSE models, and it can obtain more accurate analysis results of multimodal sentiments. This is in the process of modal feature extraction to model single modal low-level features, which has a significant advantage in capturing contextual relationships, and thus, can obtain richer high-level feature information. In addition, a local-global feature fusion module was designed, which combines local modal related features and global modal unique features to achieve comprehensive judgment of sentiment.

Comparative on the MELD

In Figure 5 and Table 6, when the CMU-MOSEI dataset is used, Acc and F1 Score of the model reach 0.8727 and 0.8600, respectively. Corr is higher than TFN and slightly lower than UniMSE. The minimum MAE is only 0.5011. The data fully show that this model has better effect than TFN and UniMSE models on MELD dataset and can obtain more accurate multi-modal sentiment analysis results. This is because the model uses a hierarchical structure to process multimodal data, including audio, text, images, and so on. By feature extraction and emotion classification for each mode separately, the model can better learn the correlation and complementarity between different

Figure 4. Results in the CMU-MOSEI

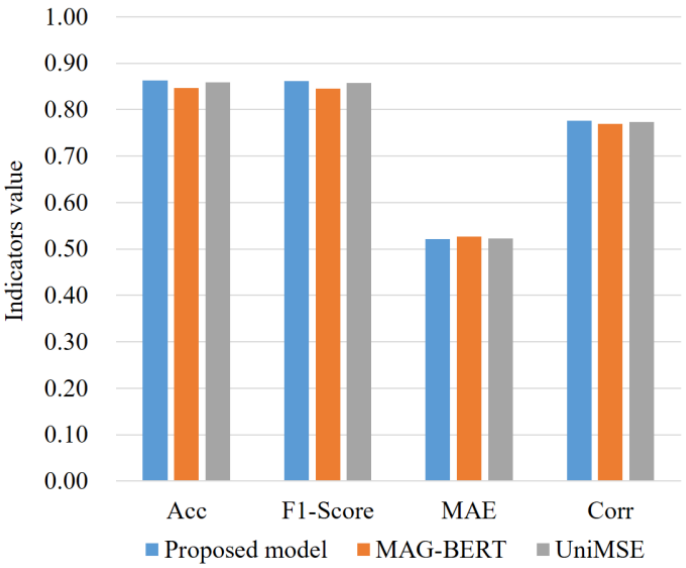


Table 5. Results in the CMU-MOSEI

Model Indicator	Proposed Model	MAG-BERT (Rahman, W., et al., 2020)	UniMSE (Hu, G., et al., 2022)	TFEE(Le, H. D., et al., 2023).	MLCCT(Gong, P., et al,. 2023)
Acc	0.8627	0.8470	0.8586	0.6781	0.632
F1-Score	0.8610	0.8450	0.8579	0.4760	0.624
MAE	0.5210	0.5260	0.5230	-	-
Corr	0.7760	0.7690	0.7730	-	-

Figure 5. Results in the MELD

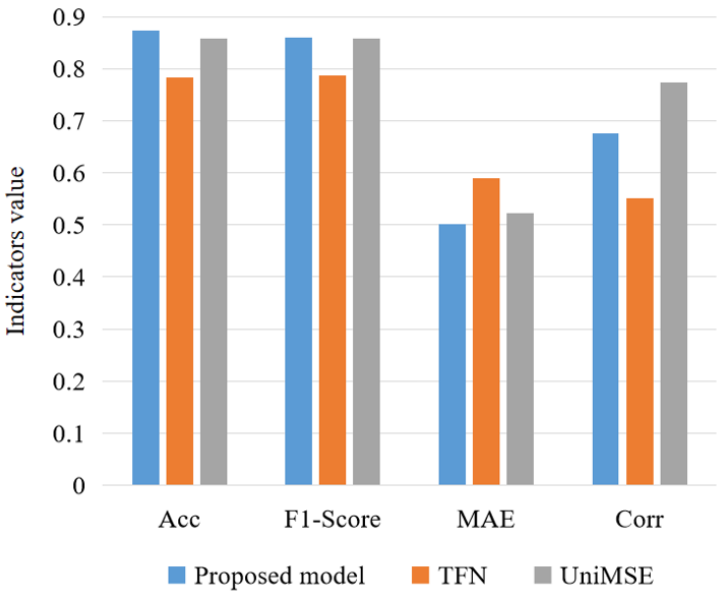


Table 6. Results in the MELD

Model Indicator	Proposed Model	TFN (Zadeh, A., et al., 2017)	UniMSE (Hu, G., et al., 2022)
Acc	0.8727	0.7838	0.8586
F1-Score	0.8600	0.7862	0.8579
MAE	0.5011	0.5903	0.5230
Corr	0.6760	0.5518	0.7730

modes. Secondly, the hierarchical adaptive feature fusion network model also introduces an adaptive mechanism, which can automatically select and weigh the importance of modes according to specific tasks and data, and further improve the interpretation ability of the model. Through adaptive feature fusion, the model can adjust the modal weights according to the characteristics of the data and the contribution degree of the modes, so as to better control the influence of different modes in the data. Additionally, it has a more effective feature extraction method, so that it has better performance in multi-modal sentiment analysis task.

Comparative on the CH-SIMS Dataset

The following is a comparative of the proposed model with TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), and ConFEDE (Yang et al., 2023) based on the CH-SIMS dataset:

(1) TFN (Zadeh et al., 2017): First, create multidimensional tensors to represent various modal features, and then dynamically exchange information between modalities through external product calculation.
(2) MulT (Tsai et al., 2019): A cross-channel attention interaction module was designed using the structure, which focuses on multi-channel sequence interactions with different time steps.
(3) MISA (Hazarika et al., 2020): Maps each modality to modal private space and cross-modal shared space to achieve the integration of interactive information within and between modalities.
(4) Self-MM (Yu et al., 2021): Generate unimodal labels through a designed self-supervised learning strategy, and train both unimodal and jointly to learn consistency and differences between modalities.
(5) ConFEDE (Yang et al., 2023): Interactions between modalities are achieved by modeling specific view interactions and cross-view interactions, and they are fused in the temporal dimension through multi-view gating mechanisms.

Under the same experimental conditions, different evaluation index values calculated by different models are shown in Figure 6 and Table 7.

In Figure 6 and Table 7, when using the CH-SIMS dataset, the proposed model has the largest Acc, F1 Score, and Corr, reaching 0.8234, 0.8218, and 0.6680, respectively, while the MAE is the smallest, at 0.3902. This means that the proposed model has better performance compared to TFN, MulT, MISA, Self MM, and ConFEDE models, and can obtain more accurate analysis results of sentiment. This is because the introduction of a local cross-modal interaction module effectively improves the problem of insufficient information fusion between modalities speech and visual modalities with relatively small contributions as auxiliary modalities and using cross-modal attention to achieve important

Figure 6. Results in the CMU-MOSEI

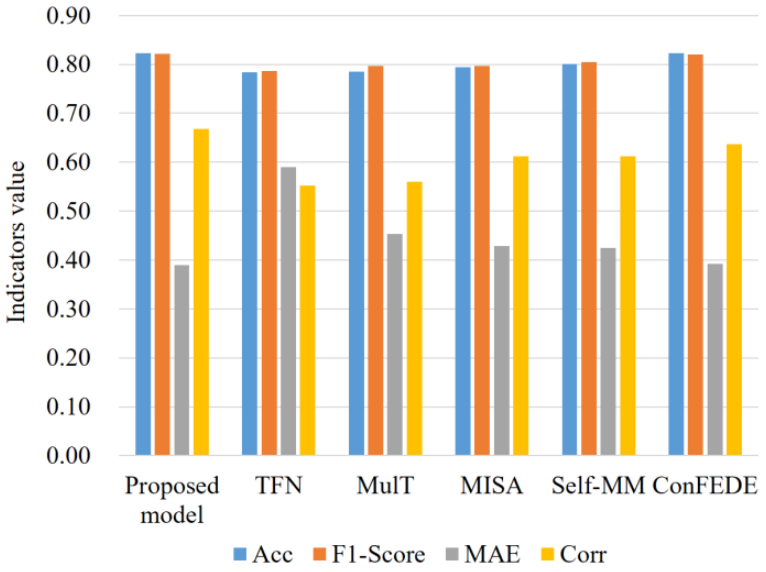


Table 7. Results in the CMU-MOSEI

Indicator Model	Acc	F1-Score	MAE	Corr
Proposed model	0.8234	0.8218	0.3902	0.6680
TFN (Zadeh, A., et al., 2017)	0.7838	0.7862	0.5903	0.5518
MulT (Tsai, Y.H., et al., 2019)	0.7856	0.7966	0.4530	0.5594
MISA (Hazarika, D., et al., 2020)	0.7943	0.7970	0.4285	0.6123
Self-MM (Yu, W., et al., 2021)	0.8004	0.8044	0.4252	0.6118
ConFEDE (Yang, J., et al., 2023)	0.8223	0.8208	0.3920	0.6370

information representation between the two modalities. In addition, the introduction of the global multimodal feature interaction module can achieve hierarchical adaptive fusion based on important information through the adaptive gating mechanism, greatly improving the accuracy of the model.

Ablation

Modal

To further illustrate the importance and differences of each part, ablation experiments will be conducted on the proposed model. Taking the CH-SIMS dataset as an example, simulation analysis was conducted on models containing A+V, A+T, T+V, and A+V+T. The simulation results are shown in Table 8.

In Table 8, A+V represents the model that only includes audio and video modal extraction. A+L represents a model that only includes audio and text modal extraction. L+V represents a model that only includes text and video modal extraction. A+V+L represents a model that includes text, audio, and video modal extraction. According to the results of the ablation experiment, the "L+V" model are superior to the "A+V" and "L+A" models, while the "L+A" model are superior to the "A+V" model. This indicates that text information plays the greatest role in the multimodal feature extraction process, followed by video information, and finally, audio information. However, the "A+V+L" model can only exhibit the best MSA performance when the three are extracted and fused.

This is because the introduction of a local cross-modal interaction module effectively improves the problem of insufficient information fusion between modalities by using speech and visual modalities with relatively small contributions as auxiliary modalities and using cross-modal attention to achieve important information representation between the two modalities. In addition, the introduction of the global feature interaction module can achieve hierarchical adaptive fusion based on important information through the adaptive gating mechanism, greatly improving the accuracy of the model.

Modal

To verify the different importance of different modalities on the final MSA results, the model conducted MSA experiments using text mode, speech mode, and visual mode as guidance modes and compared the experimental results. The results in CH-SIMS are as follows.

Table 8. Modal ablation experimental results

Indicator Model	Acc	F1-Score	MAE	Corr
L+V	0.7057	0.7043	0.4544	0.5725
A+V	0.6530	0.6517	0.5094	0.5297
L+A	0.6785	0.6772	0.4715	0.5504
A+V+L	0.8234	0.8218	0.3902	0.6680

In Table 9, the proposed model exhibits the best performance when using text mode as the guidance mode, speech mode, and visual mode as auxiliary modes. This further indicates that text modality plays an important role in the emotional judgment of the final data. When using speech mode or visual mode as the guidance mode, the accuracy, F1 Score value, and Corr of MSA all showed a significant decrease, while MAE increased. This indicates that different modalities have different degrees of importance for the final MSA results in MSA tasks. The greatest contribution of text modality to MSA results reflects the importance of text modality.

Model

To further validate the effectiveness of each module, a model ablation experiment was designed to compare the impact of different modules. The design is as follows:

Model 1: w/o Multi-head self-attention (MHSA): Remove the high-level feature extraction module based on MHSA from the complete model, and directly concatenate and adaptively fuse the low-level features.
Model 2: w/o Cross attention: Based on the complete model, the improved cross-modal attention module is removed, and after self-attention, the single modal vectors are directly concatenated for splicing and adaptive fusion.
Model 3: w/o adaptive gating unit: Remove the fusion module based on the adaptive gating unit from the complete model, and directly stack and fully connect the fused high-level fusion features.
Model 4: w/o text gate, w/o speech gate, and w/o visual gate: In the fusion process based on the adaptive gating mechanism, the text gate, speech gate, and visual gate are sequentially removed.
Model 5: w/o related features: Remove modal-related features in the local global feature fusion module and only use modal-specific features.
Model 6: w/o unique features: Remove modal unique features in the local global feature fusion module and only use modal-related features.

The experiment was conducted using the CH-SIMS dataset, and the final model ablation experiment results are as follows.

In Table 10, the results on the CH-SIMS showed that removing any module from the model resulted in a decrease in Acc, F1 Score, and Corr of the model SA, as well as an increase in MAE. Therefore, removing modal-related or modal-specific features will affect the overall performance of the model. When the three features are fused simultaneously, the proposed model can learn more feature information, which is more conducive to SA. At the same time, this fully verifies the necessity of each module for the proposed model to achieve the best experimental results.

CONCLUSION

A hierarchical adaptive feature fusion network-based MSA method is proposed to address the issue of insufficient utilization of modal importance information caused by the current MSA method's focus on multimodal feature fusion, resulting in the loss of important information in the modality. The

Table 9. Modal importance ablation experimental results

Indicator Model	Acc	F1-Score	MAE	Corr
V+L+A (Visual attention)	0.8143	0.8128	0.4159	0.6607
A+V+L (Audio attention)	0.8086	0.8070	0.4232	0.6560
L+A+V (Text attention)	0.8234	0.8218	0.3902	0.6680

Table 10. Model

Indicator Model	Acc	F1-Score	MAE	Corr
Model 1	0.8110	0.8095	0.4043	0.6580
Model 2	0.8069	0.8054	0.4124	0.6546
Model 3	0.8135	0.8119	0.4055	0.6600
Model 4	0.8086	0.8070	0.4132	0.6560
Model 5	0.8053	0.8037	0.4116	0.6533
Model 6	0.8119	0.8103	0.4047	0.6586
Proposed model	0.8234	0.8218	0.3902	0.6680

performance of this method was verified through experiments. The results indicate that extracting features from text, video, and audio modalities based on RoBERTa, ResViT, and LibROSA can improve the effectiveness of information feature extraction. By combining the multimodal feature extraction module and the cross-modal feature interaction module, the model can effectively learn information and achieve an interactive fusion of information between modalities. The introduction of an adaptive gate control mechanism can effectively improve the global multimodal feature interaction process and improve the accuracy of MSA.

However, this work also has some limitations, such as the lack of consideration for the dual recognition of sentiments and emotions, and the generalization ability of the proposed model needs to be verified. Future work will further explore the semantic interaction between textual and non-textual modalities in MSA, as well as the dual recognition of sentiments. In addition, based on this, focus is on studying the impact of different sentiment information contained in different scenes on the emotional expression of characters, to further enhance the generalization of the MSA model.

AUTHOR NOTE

Huchao Zhang: https://orcid.org/0009-0005-0703-3408

We have no known conflict of interest to disclose.

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

Correspondence concerning this article should be addressed to Zhejiang Agricultural Business College, Shaoxing, Zhejiang, 312088, China.

Corresponding Author: Huchao Zhang, Email: 18158653170@163.com

REFERENCES

Ahmed, S., Rajput, A., Sarirete, A., & Chowdhry, T. J. (2022). Flesch-Kincaid measure as proxy of socio-economic status on Twitter: Comparing US senator writing to Internet users. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–19. doi:10.4018/IJSWIS.297037

Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., Gupta, B. B., Gupta, B. B., & Gupta, B. B. (2022). Phishing website detection with semantic features based on machine learning classifiers: A comparative study. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–24. doi:10.4018/IJSWIS.297032

Balcilar, M., Bouri, E., Gupta, R., & Kyei, C. K. (2021). High-frequency predictability of housing market movements of the United States: The role of economic sentiment. *Journal of Behavioral Finance*, *22*(4), 490–498. doi:10.1080/15427560.2020.1822359

Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K. (2022). K-Means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, *70*(1), 1016–1024. doi:10.1109/TIE.2022.3150097

Gao, S. (2022). A two-channel attention mechanism-based mobileNetV2 and bidirectional long short memory network for multi-modal dimension dance emotion recognition. *Journal of Applied Science and Engineering*, *26*(4), 455–464.

Garcia-Garcia, J. M., Lozano, M. D., Penichet, V. M., & Law, E. L. C. (2023). Building a three-level multimodal emotion recognition framework. *Multimedia Tools and Applications*, *82*(1), 239–269. doi:10.1007/s11042-022-13254-8

Gong, P., Liu, J., Wu, Z., Han, B., Ken Wang, Y., & He, H. (2023). A multi-level circulant cross-modal transformer for multimodal speech emotion recognition. *Computers, Materials & Continua*, *74*(2), 4204–4220. doi:10.32604/cmc.2023.028291

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., & Hu, S. M. (2022). Attentions in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. doi:10.1007/s41095-022-0271-y

Hazarika, D., Zimmermann, R., & Poria, S. (2020). Misa: Modality-invariant and specific representations for multimodal sentiment analysis. *Proceedings of the 28th ACM International Conference on Multimedia*, 1122-1131. doi:10.1145/3394171.3413678

Hu, G., Lin, T., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022). UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *Conference on Empirical Methods in Natural Language Processing*, 1-15 doi:10.18653/v1/2022.emnlp-main.534

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. *International Journal of Multimedia Information Retrieval*, *9*(5), 103–112. doi:10.1007/s13735-019-00185-8

Kamath, P. S., & Kim, R. W. (2007). The model for end-stage liver disease (MELD). *Hepatology (Baltimore, Md.)*, *45*(3), 797–805. doi:10.1002/hep.21563 PMID:17326206

Kumar, A., & Sivakumar, P. (2022). Cat-squirrel optimization algorithm for VM migration in a cloud computing platform. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–23.

Le, H. D., Lee, G. S., Kim, S. H., Kim, S., & Yang, H. J. (2023). Multi-Label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access : Practical Innovations, Open Solutions*, *16*(11), 14742–14751. doi:10.1109/ACCESS.2023.3244390

Liao, W., Zeng, B., Liu, J., Wei, P., & Fang, J. (2022). Image-text interaction graph neural network for image-text sentiment analysis. *Applied Intelligence*, *52*(10), 11184–11198. doi:10.1007/s10489-021-02936-9

Mai, S., Hu, H., & Xing, S. (2020). Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 164–172. doi:10.1609/aaai.v34i01.5347

Mohammed, S. S., Menaouer, B., Zohra, A. F. F., & Nada, M. (2022). Sentiment analysis of COVID-19 tweets using adaptive neuro-fuzzy inference system models. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–20. doi:10.4018/IJSSCI.300361

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention of deep learning. *Neurocomputing*, *452*(3), 48–62. doi:10.1016/j.neucom.2021.03.091

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2023). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, *14*(1), 108–132. doi:10.1109/TAFFC.2020.3038167

Rahman, W., Hasan, M., Lee, S., Zadeh, A., Mao, C., Morency, L., & Hoque, E. (2020). Integrating multimodal information in large pertained transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359-2369. doi:10.18653/v1/2020.acl-main.214

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983. doi:10.1016/j.asoc.2020.106983

Salhi, D. E., Tari, A., & Kechadi, M. T. (2021). Using e-reputation for sentiment analysis: Twitter as a case study. *International Journal of Cloud Applications and Computing*, *11*(2), 32–47. doi:10.4018/IJCAC.2021040103

Schneider, C. R., & van der Linden, S. (2023). An emotional road to sustainability: How affective science can support pro-climate action. *Emotion Review*, *15*(4), 284–288. doi:10.1177/17540739231193742

Singh, S. K., & Sachan, M. K. (2021). Classification of code-mixed bilingual phonetic text using sentiment analysis. *International Journal on Semantic Web and Information Systems*, *17*(2), 59–78. doi:10.4018/IJSWIS.2021040104

Su, C. W., Liu, Y., Chang, T., & Umar, M. (2023). Can gold hedge the risk of fear sentiments? *Technological and Economic Development of Economy*, *29*(1), 23–44. doi:10.3846/tede.2022.17302

Sun, Z., Sarma, P., Sethares, W., & Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *3*(5), 8992–8999. doi:10.1609/aaai.v34i05.6431

Tiwari, A. K., Bathia, D., Bouri, E., & Gupta, R. (2021). Investor sentiment connectedness: Evidence from linear and nonlinear causality approaches. *Annals of Financial Economics*, *16*(04), 2150016. doi:10.1142/S2010495221500160

Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558-6569. doi:10.18653/v1/P19-1656

Wang, F., Tian, S., Yu, L., Liu, J., Wang, J., Li, K., & Wang, Y. (2022). TEDT: Transformer-Based encoding–decoding translation network for multimodal sentiment analysis. *Cognitive Computation*, *3*(12), 1–15.

Wang, Z., Wan, Z., & Wan, X. (2020). Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. *Proceedings of The Web Conference 2020*, 2514-2520. doi:10.1145/3366423.3380000

Xi, C., Lu, G., & Yan, J. (2020). Multimodal sentiment analysis based on multi-head Attention. *Proceedings of the 4th international conference on machine learning and soft computing*, 34-39. doi:10.1145/3380688.3380693

Yadav, A., & Vishwakarma, D. K. (2023). A deep multi-level attentive network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing Communications and Applications*, *19*(1), 1–19. doi:10.1145/3517139

Yang, B., Shao, B., Wu, L., & Lin, X. (2022). Multimodal sentiment analysis with unidirectional modality translation. *Neurocomputing*, *467*(3), 130–137. doi:10.1016/j.neucom.2021.09.041

Yang, J., Yu, Niu, D., Guo, W., & Xu, Y. (2023). ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 7617–7630. doi:10.18653/v1/2023.acl-long.421

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 AAAI conference on artificial intelligence,* 10790-10797. doi:10.1609/aaai.v35i12.17289

Yuan, Z., Li, W., Xu, H., & Yu, W. (2021). Transformer-based feature reconstruction network for robust multimodal sentiment analysis, *Proceedings of the 29th ACM International Conference on Multimedia, 4*(6), 4400-4407. doi:10.1145/3474085.3475585

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103-1114. doi:10.18653/v1/D17-1115

Zhang, F., Li, X., Lim, C. P., Hua, Q., Dong, C., & Zhai, J. (2022). Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. *Information Fusion*, *88*(1), 296–304. doi:10.1016/j.inffus.2022.07.006

Zhang, J., Liu, X., Chen, M., Ye, Q., & Wang, Z. (2022). Image sentiment classification via multi-level sentiment region correlation analysis. *Neurocomputing*, *469*(2), 221–233. doi:10.1016/j.neucom.2021.10.062

Zhang, Q., Shi, L., & Liu, P. (2022). ICDN: Integrating consistency and difference networks by transformer for multimodal sentiment analysis. *Applied Intelligence*, *8*(5), 1–14. PMID:36531970

APPENDIX

Table 11. Appendix

F_{kl}^{L}	text feature
F_{kl}^{A}	speech feature
F_{kl}^{V}	video feature
H_k	discourse amount contained in the *k*-th video
D_k^n	feature dimensions of each modality in the *k*-th video
F_k^L	*k*-th video into transformer
ω_Q	weight matrices for text features
ω_K	weight matrices for text features
ω_V	weight matrices for text features
$\lambda_k^L = \lambda_k = \lambda_v$	corresponding dimension size
$\lambda_k^h = \lambda_v^h = \frac{\lambda_k^L}{h}$	dimensional size of each head.
H_L	high-level text feature
λ_L	feature dimension
H_A	high-level video feature representation
l_A	the sequence lengths of audio
l_V	the sequence lengths of video
λ_A	feature dimensions of audio
λ_V	feature dimensions of video
H_m	local correlation features of the three modes
t_A	similarity weight between multimodal related features and speech features
H_L^*	text specific features

continued on following page

Table 11. Continued

H_A^*	speech specific features
H_V^*	visual specific features
Y	final prediction result
H_m'	modal related feature
H'^*_m	modal specific feature

Huchao Zhang was born in Shaoxing, Zhejiang. P.R. China, in 1988. He received the master's degree from Jiangxi Normal University, P.R. China. Now, he works in he works in department of Basic Education, Zhejiang Agricultural Business College. His research interest include intelligent educational software, artificial intelligence.