# Cross-Modal Sentiment Analysis on Social Media using Improved Nonverbal Representation Learning and GHRNN Fusion

Minglun Xue[1*], Dongyang Wang[2]
[1]Department of Primary Education, Jiaozuo Normal College, Jiaozuo 454000, China
[2]Department of Foreign Languages, Shangqiu Normal University, Shangqiu 476000, China
E-mail: minglunxue@outlook.com
[*]Corresponding author's

**Keywords:** social media, multi-modal data, information extraction, fusion network, representation learning

**Received:** May 6, 2025

*Traditional sentiment analysis methods mainly focus on textual data, while human emotions are multidimensional, usually related to sound, body language, etc. The use of multi-modal data can provide a deeper and broader understanding of human sentiment conveyance. To address the above issues, a method for extracting and analyzing emotional information features based on improved nonverbal representation learning networks and multi-modal data (Improved Nonverbal Representation Learning Networks and MD, INPNRLN-MD) is proposed. On this basis, an improved multi-modal data fusion method based on Gated Hierarchical Recurrent Neural Network and Cross-Modal Attention (GHRNN-CMA) is designed for the MD fusion part. Compared with traditional baselines, INPNRLN-MD extracts text features through the BERT model and utilizes ELN to process audio and video data, which can more effectively capture emotional information in multi-modal data. The cross-modal attention mechanism of GHRNN-CMA can enhance the interaction between modalities and improve the accuracy of emotion information recognition. Finally, the performance of the model is validated on the CMU-OSI and CMU-MCSEI datasets using indicators such as F1 value, Pearson correlation, mean absolute error, and second-order accuracy. During the training process, the study used a single NVIDIA GTX TITAN X GPU for testing, with 12 VRAM and a batch size of only 32 to converge. The inference stage has a relatively light computational load and can be deployed on ordinary cloud servers or edge devices. The research results show that compared with mainstream algorithms, the emotion information feature extraction and analysis method based on improved nonverbal representation learning network and multi-modal data performs the best, with F1 score, Pearson correlation, average absolute error, and second-order accuracy reaching 83.06/85.12, 0.803, 0.696, and 83.17/85.23, respectively. The average absolute error, Pearson correlation, F1 score, and second-order accuracy of the improved multi-modal data fusion method have been improved by 1.0%, 14.67%, 3.1%, and 3.3% respectively compared to the latest method. The above results indicate that research methods are helpful in perceiving and analyzing human emotions, which is beneficial for understanding and predicting human behavior in the future, and is of great significance for maintaining social relationships and improving social governance.*

*Povzetek: Metoda INPNRLN-MD z BERT in ELN ter fuzija GHRNN-CMA omogočata bolj kvalitetno čezmodalno analizo sentimenta na družbenih omrežjih, saj presegata obstoječe pristope v točnosti in robustnosti.*

## 1 Introduction

Social media has become an indispensable part of daily life, with billions of users sharing information, interacting, and communicating on the platform every day [1]. These massive amounts of data contain useful insights that can aid in comprehending user behavior, analyze trends, and improve products and services [2]. Sentiment analysis is a core task in natural language processing, aimed at identifying emotional states of opinions, emotions, and evaluations [3]. Due to the diversity of social media information, it is of great significance to comprehensively consider text and image information for multi-modal sentiment analysis in order to improve the accuracy of sentiment analysis on social media [4]. In addition, different multi-modal data (MD) from social media can provide very important clues and are vital in enhancing the recognition and detection performance of subtasks [5-8]. Therefore, compared with single-modal analysis, multi-modal analysis of social media data effectively utilizes the relationship and influence between visual and textual information, which not only helps scholars to accurately understand people's attitudes and habits towards life in the real world, but also enables them to grasp people's choices in areas such as healthcare, political topics, TV movies, and online shopping [9]. The research questions are as follows: can specific modalities of nonverbal learning improve emotion classification, and can emotion information feature extraction and analysis methods based on improved nonverbal representation learning networks and MD more accurately perceive and analyze human emotions. To solve the above problems, a method for extracting and analyzing emotional information features based on improved nonverbal representation learning network and MD (INPNRLN-MD) is proposed. On this basis, an improved MD fusion method based on gated hierarchical recurrent neural network and cross-modal attention (GHRNN-CMA) is designed for the MD fusion part. The research objective is to maintain the accuracy of A7 while reducing the average absolute error and to enhance the overall performance of sentiment analysis. The innovation of the research lies in two aspects. On the one hand, it proposes an emotion information feature extraction and analysis method based on INPNRLN-MD to achieve faster understanding of the emotions and viewpoints that users want to express and share. On the other hand, it designs an improved MD fusion method based on GHRNN-CMA to fully utilize the information differences between different modalities and obtain the optimal feature representation.

## 2 Related works

In today's rapidly developing world of artificial intelligence, sentiment analysis has become a key tool for understanding human emotions and attitudes. However, traditional sentiment analysis methods are often limited to a single data modality, which limits their comprehensiveness and accuracy. To gain a deeper understanding of human emotions, it is necessary to introduce MD fusion to raise the precision and breadth of sentiment analysis. Numerous scholars have conducted in-depth analysis and exploration on this matter. Lei et al. proposed three sets of paired sorting rules, namely support vector machine, deep neural network, and gradient enhanced decision tree, to generate multi-modal sentiment information analysis for preference learning. The research results showed that compared with the traditional classifier baseline, the performance of the three preference learning models used in the study was better, and the model with gradient enhanced decision tree had the best performance. In addition, the results of the multi-modal emotion dataset annotated by actor performance crowd-sourcing showed that by combining the two best ranking models through research methods, the optimal overall accuracy was 85.06% [10]. Zhu et al. analyzed different multi-modal sentiment analysis techniques, including machine learning and deep learning methods, involving fusion strategies for different modalities. The results showed that multi-modal sentiment analysis techniques made progress in understanding human emotional expression, but still faced challenges in modal fusion and context understanding [11]. Chen et al. proposed a method for processing MD grounded on K-Meand and kernel canonical correlation analysis. The experiment outcomes indicated that the research method validly improved the heterogeneity between different modalities, promoted multi-modal emotion recognition, and greatly improved the recognition rate. Its performance on the Sari audio-visual expression emotion dataset was 2.77% higher than the average level, and on the enterface'05 audio-visual emotion dataset it was 4.7% higher [12]. Mai et al. designed a new framework for three mode mixed comparative information to fully explore cross-modal interaction and reduce modal gap. Numerous experiments showed that the research method was superior to the baseline in multi-modal sentiment analysis and emotion recognition [13]. Chiorrini et al. focused on a large amount of unstructured data online, which contains users' emotions and feelings towards various topics that cannot be achieved through traditional search engines. Therefore, they developed a new architecture for emotion aware search engines and used deep learning-based emotion recognition algorithms to extract emotion vectors. The results confirmed the superiority of this method [14]. To explore how to mine the required information from massive Internet data, Fang et al. designed a method based on convolutional neural network and time convolutional network, and added multi-modal attention mechanism and cross-modal transformer structure on this basis. The findings indicated that the precision of this technique was 2.88% higher than that of using time convolutional network alone [15]. Gupta et al. aimed to investigate emotional disorders through emoticons, and therefore proposed a multi-modal emotion recognition method based on a multi-view ensemble learning model. The outcomes indicated that the method provided an accuracy of 88.29 [16]. The summary table of the relevant works mentioned above is shown in Table 1.

Table 1: Summary table of related works

| Literature | Method | Data set | Key performance | Limitation |
|---|---|---|---|---|
| [10] | Multi model ensemble based on preference learning | Actor performance crowdsourcing annotation dataset | The optimal combination accuracy is 85.06% | Unresolved modal heterogeneity issues and lack of dynamic interaction modeling |
| [11] | Overview of Multi-modal Technologies | - | Pointing out that modal fusion and contextual understanding are the main challenges | Unprocessed non aligned sequence issues |
| [12] | K-Means+nuclear canonical correlation analysis | urrey AVED、enterface'05 | The recognition rate has increased by 2.77% and 4.7%, respectively | Unmodeled high-order nonlinear modal interactions, ignoring |

| | | | | |
|---|---|---|---|---|
| | | | | temporal dynamics |
| [13] | Three mode hybrid comparative learning framework | CMU-MOSI、CMU-MCSEI | Better than baseline | Explicitly processing modal redundancy information |
| [14] | DL based emotion perception search engine | Unstructured data on the internet | Better than baseline | Unsolved problem of imbalanced MD fusion |
| [15] | CNN+TCN+Multi-modal Attention | E-commerce product marketing data | Compared to TCN, the accuracy has increased by 2.88% | Attention mechanism does not differentiate modal contributions |
| [16] | Multi view ensemble learning model | Social media data containing emoticons | The accuracy rate is 88.29% | Unprocessed modal missing scenarios |

Based on the above content, it can be concluded that current research mainly focuses on using MD for sentiment analysis and applying it in various fields. However, existing methods have significant shortcomings in non-verbal representation optimization, modal dynamic fusion, and cross-modal interaction modeling. The design of INPNRLN-MD and GHRNN-CMA directly addresses these shortcomings and can effectively compensate for the shortcomings of existing research. Therefore, the study proposed an emotion information feature extraction and analysis method based on INPNRLN-MD and an improved MD fusion method based on GHRNN-CMA.

# 3 Emotional information extraction and analysis for MD on social media

To fully utilize the massive emotional information of social media MD, a feature extraction method based on INPNRLN-MD was studied and designed, and an improved MD fusion method based on GHRNN-CMA was proposed to improve the fusion part. INPNRLN-MD is a component for feature extraction and preliminary analysis, which is part of the GHRNN-CMA method.

## 3.1 Emotional information feature extraction and analysis method based on INPNRLN-MD

Social media, as an important platform for information exchange in people's daily lives, has accumulated a massive amount of user generated content in recent years. The information forms are diverse, including text, audio, images, and other data forms [17,18]. Leveraging MD to fully grasp users' emotional expressions and enhance sentiment analysis performance is crucial in the realm of sentiment analysis. Seamlessly integrating intricate multi-modal interactions to uncover potentially valuable emotional features poses a significant challenge in this field. Therefore, in this section, a sentiment information feature extraction method based on INPNRLN-MD is proposed, which focuses on extracting sentiment features from different modalities such as text, audio, and video, and provides input for subsequent sentiment analysis. The MD sequence of this model mainly includes three modalities: text, audio, and video. Assuming that the modal information of text, audio, and video are represented as $t$, $a$, and $v$, respectively, and the input content can be represented by $z\in\{t,a,v\}$, where the sequence length and corresponding feature dimensions of the above content are represented as $L_{z\in\{t,a,v\}}$ and $d_{z\in\{t,a,v\}}$, respectively. The emotional information feature extraction part of the text modal data is completed using a pre-trained Bidirectional Encoder Representation from Transformers (BERT) model, as shown in Figure 1.

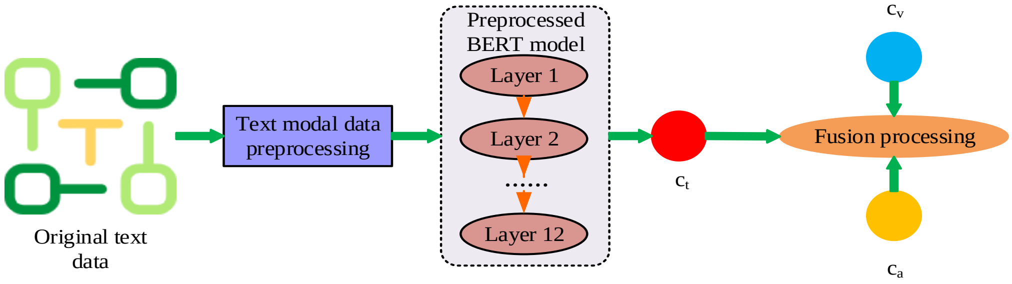


Figure 1: Schematic diagram for feature extraction of text modal emotional information

In Figure 1, the input text modality data is fed into a preprocessed BERT model to obtain the feature representation $c_t$ in the text modality. It is then processed by outer product with the feature representations obtained from other modality data to obtain the final shareable feature space. The interactive representation is then obtained through contrastive learning and transmitted to the fully connected layer for sentiment prediction. The BERT model used in the study is stacked through 12 layers of Transformers and introduces a special label, the classification label $[CLS]$, whose last hidden state serves as the aggregated representation of the entire sequence; The separator mark $[SEP]$ is used to separate different sentences, marking the end of the first sentence and the beginning of the second sentence [19,21]. The above data are transferred to the BERT model together to complete the word embedding, and the first word vector of the 12th layer is used as the representation, where the implied state is represented by the average value, as shown in equation (1).

$$\begin{cases} t_i = \{[CLS], word_1, word_2, \cdots, word_n, [SEP]\} \\ c_t = BERT \left(t_i, \zeta_t^{\text{BERT}}\right) \in R^{d_t}, i \in [1, n] \end{cases} \quad (1)$$

In equation (1), $word_n$ and $n$ represent the words and their quantities in the discourse, $\zeta_t^{\text{BERT}}$ represents the hyperparameters required for the BERT model, and $R^{d_t}$ represents the space of the output vector. After the emotional information extraction of text modal data is completed, the emotional information feature extraction of audio modal data can be carried out, as shown in Figure 2.

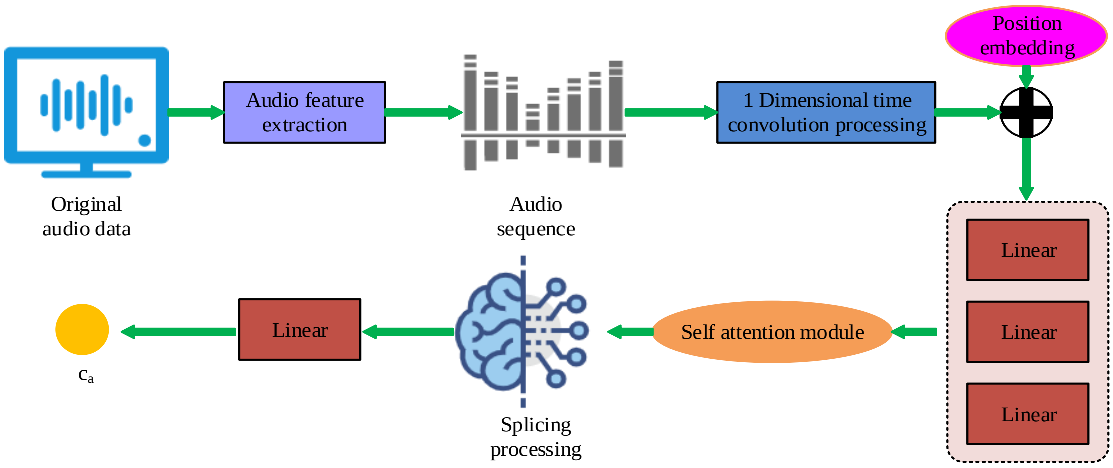


Figure 2: Schematic diagram of feature extraction of emotional information in audio modality

In Figure 2, the original audio data is first transformed into acoustic features that can be used for subsequent processing through a one-dimensional temporal convolutional layer, and then input into the Expressing Learning Network (ELN). Next, position embedding vectors are added, and then processed sequentially through linear transformation and mean weighted filters. Finally, the desired audio feature representation $c_a$ is output. The emotional information feature extraction of video modal data is similar to that of audio modal data, both using ELN, but with certain differences in the feature extractor part. Firstly, the study applies a 1D time convolution layer to the input sequence for processing, in order to enhance the perceptual power of modal data. The calculation is shown in equation (2).

$$F_k = \text{Conv1}D_k\left(k, q_k\right) \in R^{T_k \times d} \quad (2)$$

In equation (2), $F_k$ represents the processed feature representation, $\text{Conv1}D_k(\square)$ represents the 1D time convolution function, and $k$ and $q_k$ represent the input sequence and convolution kernel size of the corresponding modal data. To assign time information to a sequence, it can be solved by introducing positional embedding, and then implementing a self attention module on the processed sequence, as expressed in equation (3) [22].

$$\text{A}_k(Q, K, V) = \text{softmax}\left(\frac{Q_k K_k^{TN_k}}{\sqrt{d_k}}\right) V_k, k \in \{a, v\} \quad (3)$$

In equation (3), $\text{softmax}(\square)$ is the softmax function, $Q$, $K$, and $V$ correspond to the query vector, key vector, and value vector of the self attention mechanism, respectively. $TN$ and $d_k$ correspond to the vector transpose and the dimension of the corresponding modality. Then, to calculate the heads of each attention, it is necessary to linearly project the matrix through the corresponding weight matrix. Meanwhile, it can connect the heads of all modalities to obtain a multi-head self attention mechanism, as shown in equation (4).

$$Y_k = Concat\left(h_k^1, h_k^2, \cdots, h_k^a\right) W_k^o \quad (4)$$

In equation (4), $Concat(\square)$ and $W_k^o$ represent the weight matrices used for the stitching operation and stitching processing, respectively, while $h_k^a$ represents the head of the corresponding model's $a$ th attention. From this, the feature representations $c_a$ and $c_v$ corresponding to the audio modality and video modality can be obtained. Based on the above content, a schematic diagram of extracting emotional information features from video modalities can be obtained, as shown in Figure 3.

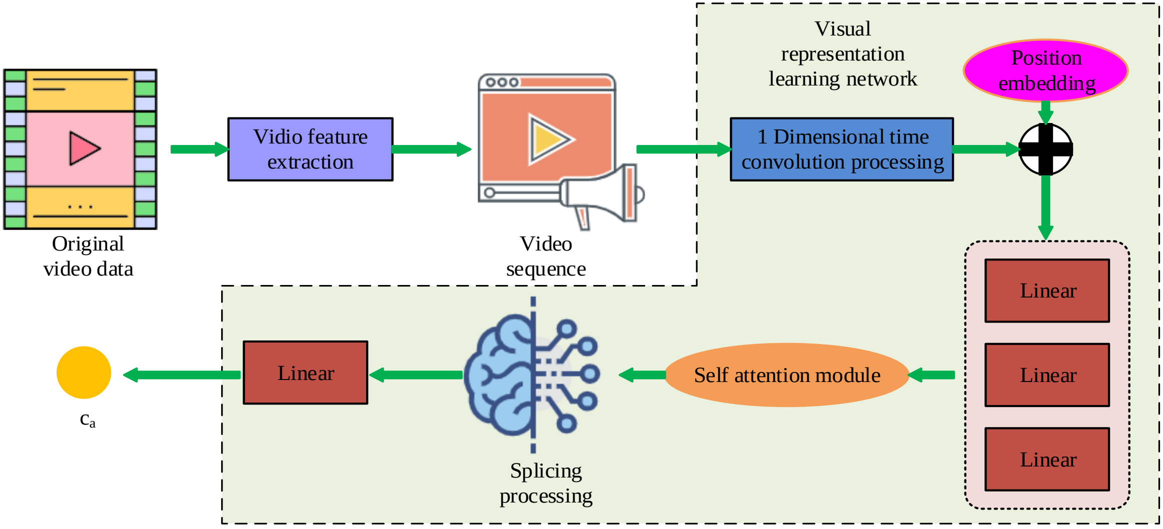


Figure 3: Schematic diagram of feature extraction of emotional information in video modality

In Figure 3, the original video modality data is processed by a 1D time convolution layer to obtain the required feature representation, which is then input into ELN for processing and position embedding. Finally, through concatenation and linear projection, the final emotional information feature representation of the video modality can be obtained. After the above data processing is completed, the outer product operation can be performed to obtain the corresponding result, as shown in equation (6).

$$\begin{cases} c_{\alpha\otimes t} = c_\alpha \otimes c_t \\ c_{v\otimes t} = c_v \otimes c_t \end{cases} \quad (6)$$

In equation (6), $c_{\alpha\otimes t}$ and $c_{v\otimes t}$ are the outer product results of the audio mode and video mode, respectively, $\otimes$ represents the dot product operation, and the dimension of the obtained vector is $m*n$. The processed MD feature representations are concatenated and projected onto $R^{d_z}$, resulting in the final output representation $c_z$. This representation can be applied in prediction, as expressed in equation (7).

$$\begin{cases} B_z = \mathrm{Re}LU\left(W_{l2}^{zT} \otimes c_z + f_{l2}^z\right), W_{l2}^{zT} \in R^{d_z\times 1} \\ e = \dfrac{1}{N}\sum_i^N \left(\left|B'_z - B_z\right|\right) \end{cases} \quad (7)$$

In equation (7), $W_{l2}^{zT}$, $f_{l2}^z$, and $\mathrm{Re}LU(\square)$ represent the processed weight matrix, bias, and modified linear unit activation function, $e$ and $N$ correspond to the average absolute error and sample size, and $B'_z$ represents the human label. In order to fully utilize the information of MD, the study introduces contrastive learning to handle the dynamics between different modalities, and constructs the loss function of the model, as shown in equation (8).

$$Loss_{total} = e' + Loss_{in} \quad (8)$$

In equation (8), $Loss_{total}$, $e'$, and $Loss_{in}$ correspond to the overall loss function, the $e$ value between the true emotion and the predicted emotion, and the interaction comparison loss between multiple modalities.

## 3.2 Improved MD fusion method based on GHRNN-CMA

Due to the unique feature dimensions and representation methods of different modalities of data, as well as differences in quantity and scale, there may be imbalances in MD fusion, and the above methods cannot fully consider the role of interaction information between modalities in emotion recognition and prediction. Therefore, the research designs an improved MD fusion method based on GHRNN-CMA, which not only includes the feature extraction part in INPNRLN-MD, but also further fuses and analyses these features through GHRNN and cross-modal attention mechanism to achieve more accurate emotion prediction. The specific process is shown in Figure 4.

In Figure 4, it mainly includes the feature extraction part based on INPNRLN-MD for corresponding modalities, GHRNN part, and cross attention part for each modality. The feature extraction part of the corresponding mode remains unchanged, while the two modal outputs and text modal representation of the outer product operation are transmitted to the GHRNN part together, which achieves the acquisition of the final predictive emotional information by removing redundant information. In addition, the different components of the model are very flexible and can be reorganized with different baselines to be applicable to various types of tasks. GHRNN schematic diagram, see Figure 5.

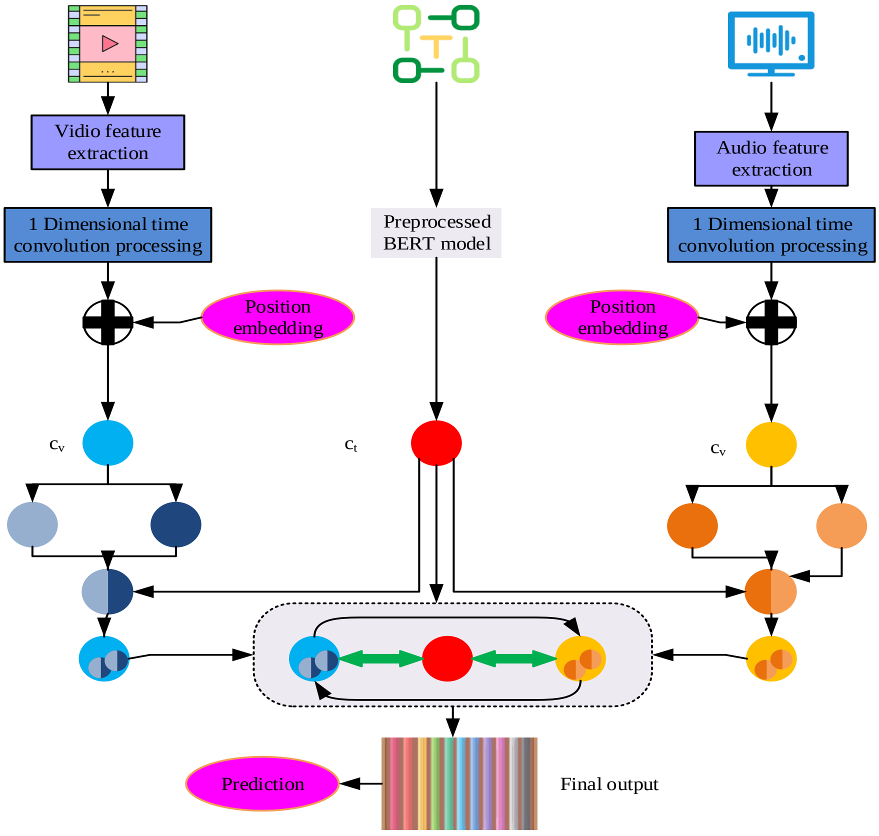


Figure 4: Flow diagram of improved MD fusion method based on GHRNN-CMA

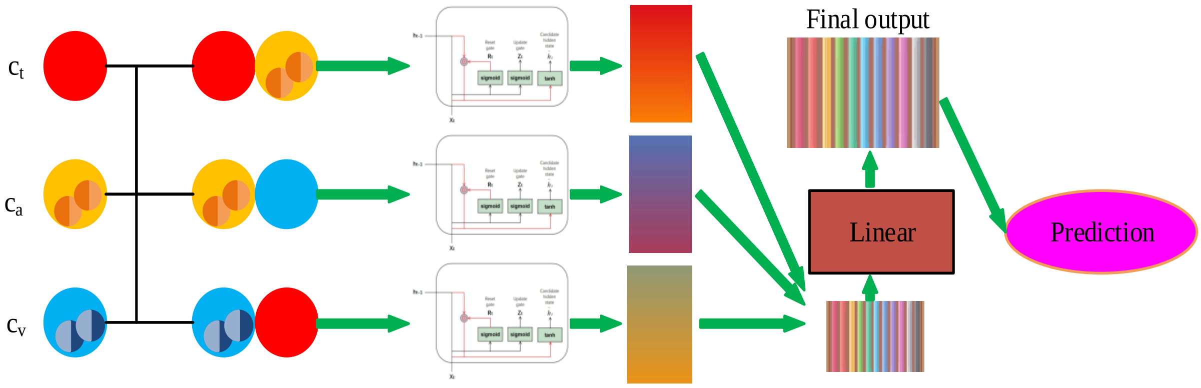


Figure 5: GHRNN schematic diagram

In Figure 5, the input text feature representation $c_t$, as well as the text-based acoustic and visual representations $c_{\alpha_t}$ and $c_{v_t}$, are processed using GHRNN. Due to the higher contribution of the text modality to sentiment information analysis compared to other modalities, the study combines the text modality feature representation with other representations to ensure the highest weight of the text modality. As a result, different representation combination expressions can be obtained, as shown in equation (9).

$$\begin{cases} c_{\alpha_t \otimes t} = Concat\left(c_t, c_{a_t}\right) \\ c_{v_t \otimes t} = Concat\left(c_t, c_{v_t}\right) \\ c_{\alpha_t \otimes v_t} = Concat\left(c_{a_t}, c_{v_t}\right) \end{cases} \quad (9)$$

In equation (9), $c_{\alpha_t \otimes t}$ and $c_{v_t \otimes t}$ represent the combination of text modal feature representation with $c_{\alpha_t}$ and $c_{v_t}$, respectively, while $c_{\alpha_t \otimes v_t}$ represents the combination of $c_{\alpha_t}$ and $c_{v_t}$. The above combination is input into GHRNN to achieve rich perception of information between various modal data, and irrelevant and redundant information is deleted. The calculation is shown in equation (10).

$$\begin{cases} c'_{\alpha t} = GHRNN\left(c_{a_t \otimes t}, \zeta'\right) \\ c'_{vt} = GHRNN\left(c_{v_t \otimes t}, \zeta'\right) \\ c'_{\alpha v} = GHRNN\left(c_{a_t \otimes v_t}, \zeta'\right) \end{cases} \quad (10)$$

In equation (10), $GHRNN\left(\bullet\right)$ and 2 represent GHRNN processing and its hyperparameters, respectively. Then transmit the output result to the fully connected layer for prediction, as expressed in equation (11).

$$\begin{cases} c_z = \mathrm{Concat}\left(c'_{at}, c'_{vt}, c'_{av}\right) \\ `c_{\ z} = \mathrm{Re}\,LU \quad \left(W_{l1}^{zT} \otimes c_z + f_{l1}^{z}\right) \end{cases} \quad (11)$$

In equation (11), $W_{l1}^{zT} \in R^{d_z \times \left(d_a + d_v + d_t\right)}$ and $f_{l1}^{z}$ represent the weights and biases processed by GHRNN, respectively. The calculation of the internal mechanism of GHRNN is shown in equation (12).

$$\begin{cases} z_t = \sigma(W_{xz} x_t + W_{hz} h_{t-1} + b_z) \\ h_t' = \tanh(W_{xh} x_t + W_{hh}(z_t \square\ h_{t-1}) + b_h) \\ h_t = (1 - z_t) \square\ h_{t-1} + z_t \square\ h_t' \end{cases} \quad (12)$$

In equation (12), $z_t$, $\sigma$, and tanh represent the gate control signal, sigmoid activation function, and hyperbolic tangent activation function, respectively. $h_t'$ and $h_t$ represent the candidate hidden state and the final hidden state, respectively. Based on the above content, the construction of an improved MD fusion method based on GHRNN-CMA can be completed. To evaluate the performance of research methods in emotional information extraction and analysis, the current mainstream F1 score, Pearson Correlation (PC), Mean Absolute Error (MAE), Accuracy 2 (A2), and Accuracy 7 (A7) were selected for analysis. The F1 score is the harmonic mean of precision and recall, used to measure the accuracy of the model in classification tasks, while the PC value is used to measure the linear relationship between two random variables, calculated by the covariance and standard deviation between the two variables, with a result range of -1 to 1. MAE is the average absolute error between predicted and observed values, and the smaller the value, the smaller the prediction error of the model. A2 is the accuracy of the model in second class classification tasks, while A7 represents the accuracy of the model in seventh class classification tasks.

# 4 Emotional information extraction and analysis results for MD

To verify the effectiveness and feasibility of the research method, the performance of the emotion information feature extraction and analysis method based on INPNRLN-MD was analyzed, and further performance comparison experiments and ablation experiments were set up to explore the effectiveness of the improved MD fusion method based on GHRNN-CMA.

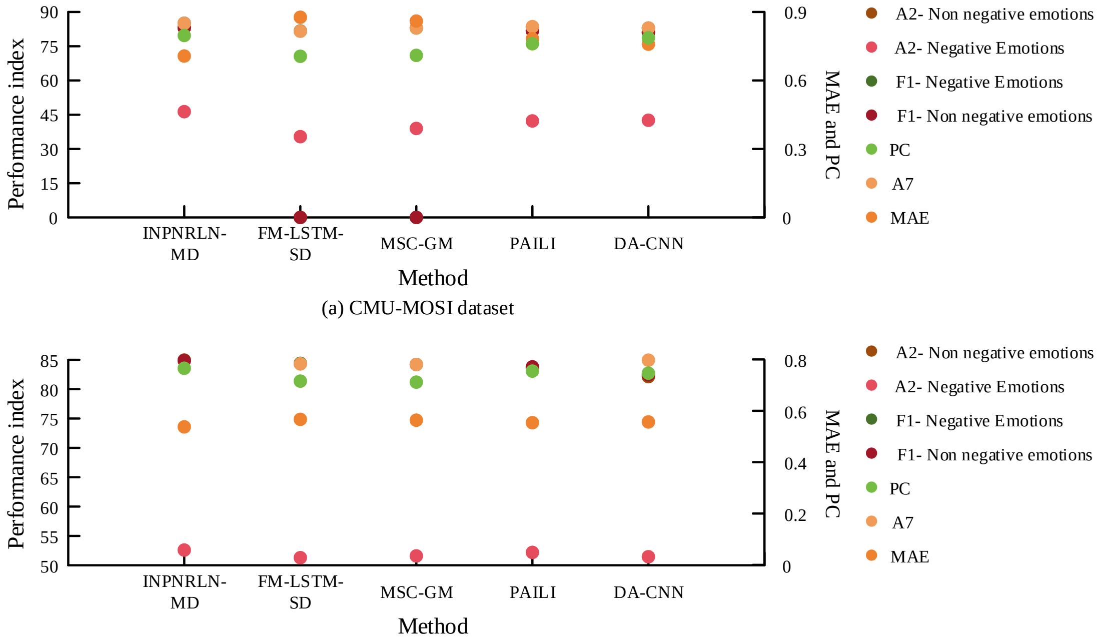
## 4.1 Results of sentiment information analysis based on INPNRLN-MD

To investigate the effectiveness of the sentiment information analysis method based on INPNRLN-MD, the experimental setup was configured as follows: a deep learning framework was employed, utilizing Pytorch 2.5.1 software, and the training process was conducted on an NVIDIA GTX TITAN X GPU. The experimental parameters were set as follows: the initial learning rates of the BERT model and other parameters were $5*10^{-5}$ and $10^{-43}$, respectively. The learning rate for audio modal data processing was set to $10^{-4}$. This study used the Adam optimizer with a weight decay coefficient of 0.01, set the iteration count to 500, batch size to 32, epochs to 30, Dropout to 0.3, and L2 regularization with a regularization coefficient of 0.00001. The experimental dataset used the multi-modal Corpus of Sentiment Intensity dataset (CMU-MOSI) and multi-modal Corpus of Sentiment and Emotion Intensity dataset (CMU-MCSEI) developed by Carnegie Mellon University (CMU). The CMU-MOSI dataset contains 2199 opinion video clips, each of which is annotated with emotional intensity. The CMU-MCSEI dataset is currently the largest multi-modal sentiment analysis and recognition dataset, containing over 23500 sentence videos from more than 1000 YouTube speakers with multi-label features. In terms of text data preprocessing, this study first removed HTML tags, special characters, and excess whitespace characters from the text, and uses natural language processing libraries such as SpaCy or NLTK for word segmentation processing. Then, the study restored the word to its stem form and removed commonly-used words. Finally, the pre trained BERT model is used to encode the text, generate word embedding representations, and fill all text sequences to the same length. In terms of audio data preprocessing, this study pre emphasized the audio signal by segmenting it into 20ms frames and normalizing the extracted features. The sample distribution and partitioning of the two datasets are shown in Table 2.

According to Table 2, the sentiment intensity of both datasets ranged from very negative to very positive, divided into 7 categories, and there were significant differences in the distribution of sentiment categories. The datasets used in the experiment were all publicly available, and their collection and use underwent corresponding ethical review and user consent procedures. In practical application scenarios, user privacy and sensitive issues were designed. Therefore, when handling social media data, it was necessary to ensure the anonymity and de identification of the data to avoid the leakage of users' personal information. Meanwhile, explicit consent from users was required to ensure that they had sufficient right to know and control over the use of their data. The specific training validation cycle is as follows: first, a pre trained BERT model is used to extract the features of the text modality and fuse them with the features of audio and video. To enhance the interaction between modalities, a multimodal data fusion method based on GHRNN-CMA is used for processing. During training, the Adam optimizer was used and the learning rate was adjusted according to experimental settings to ensure that the model can effectively learn emotional features. Each training cycle includes forward propagation, loss calculation, and backpropagation. The loss function combines cross entropy loss for sentiment classification and multimodal contrastive learning loss to ensure that the model can learn robust sentiment feature representations. To monitor model performance and prevent overfitting, the study validates the model at the end of each training cycle. The verification process includes calculating indicators such as F1 value, Pearson correlation coefficient, and mean absolute error. If the performance of the model on the validation set is better than the previous best performance, the weights of the current model are saved. At the same time, record the performance metrics on the validation set for subsequent analysis of the model's stability. If the model does not show significant performance improvement in multiple consecutive training cycles, the training should be terminated early to avoid overfitting. To verify the performance of research methods more scientifically, comparative experiments were conducted using current mainstream methods, including Fusion Matching of Long Short-erm Memory Networks and Syntactic Distance (FM-LSTM-SD), Dual Attention Convolutional Neural Network (DA-CNN), Multi-Scale Convolution and Gating Mechanism (MSC-GM), and Pay Attention to Interaction in Location Information (PAILI). To obtain fair results, the study conducted repeated validation experiments and recorded the average of twenty runs to compare the performance results of different methods on two datasets, as shown in Figure 6.

Table 2: Sample distribution and partitioning of two datasets

| Type | Emotional intensity/categorization | CMU-MCSEI | CMU-MOSI |
|---|---|---|---|
| Sample distribution of dataset | Very negative | 816 | 185 |
| | Negative | 2231 | 440 |
| | Relatively negative | 3547 | 398 |
| | Neutrality | 4998 | 96 |
| | Relatively positive | 7429 | 361 |
| | Positive | 3170 | 482 |
| | Very positive | 665 | 237 |
| Dataset partitioning | Training set | 15998 | 1542 |
| | Validation set | 2286 | 219 |
| | Test set | 4572 | 438 |



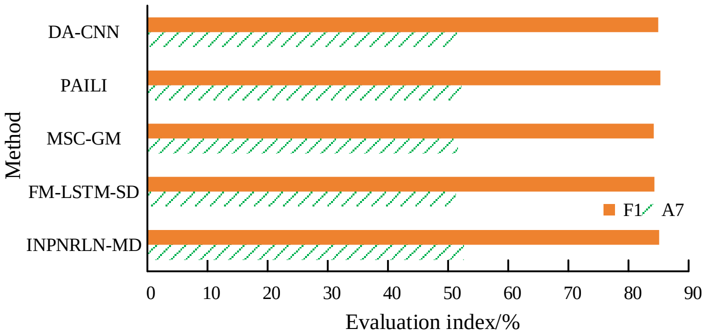
(a) CMU-MOSI dataset

(b) CMU-MOSEI dataset

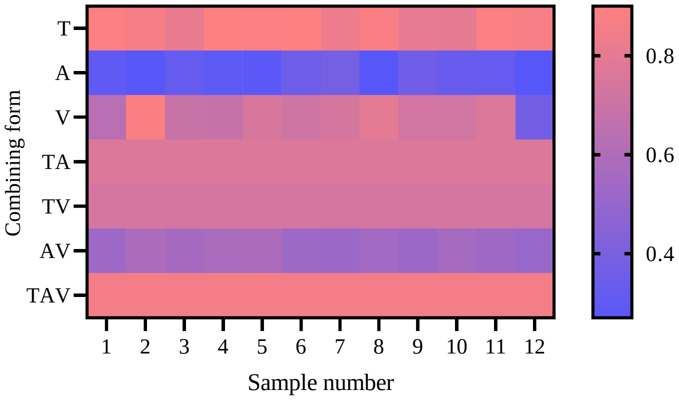
Figure 6: Comparison of performance results of different sentiment analysis methods on two datasets

Figure 6 (a) shows the performance comparison of different sentiment analysis methods on the CMU-MOSI dataset. The sentiment information feature extraction and analysis method based on INPNRLN-MD performed the best in all aspects, with F1 score, PC, MAE, A2, and A7 corresponding to 83.06/85.12, 0.803, 0.696, 83.17/85.23, and 46.41, respectively. The performance of the DA-CNN model was poor, with an MAE value as high as 0.883, a PC value of only 0.705, and an A7 value of 35.6. The MAE of the proposed INPNRLN-MD method was 0.696, which was significantly lower than the 0.883 of the DA-CNN method, indicating that INPNRLN-MD was more accurate in predicting emotional intensity. This meant in practical applications that the prediction of user emotions was closer to the true values, which helped to improve user experience and service quality. The suggested method's observed performance advantage most likely resulted from its capacity to better mimic human language patterns utilizing the CMU-MOSI dataset and completely use the underlying information in MD. In contrast, the DA-CNN model produced less than ideal results since it was less accurate at identifying important textual parts, despite being good at collecting local information. Figure 6 (b) shows the results of different sentiment analysis methods on the CMU-MCSEI dataset. The research method still exhibited excellent performance on unaligned sequences, with a significantly lower MAE value than other methods, as low as 0.527, an F1 score of 84.93/85.12, and a PC value of 0.771. To further analyze the performance of the research method, the CMU-MOSI dataset was selected for exploration, and the most critical indicators, A7 and F1 scores, were chosen. Meanwhile, 12 samples were randomly selected to analyze the impact of different modalities. The results are shown in Figure 7.

Figure 7 (a) shows the performance of different methods on important indicators. The INPNRLN-MD method outperformed other mainstream methods significantly and had the best performance on the A7 indicator, with an improvement of 28.3% compared to the FM-LSTM-SD method and an improvement of about 8.6% on the F1 score indicator. Figure 7 (b) shows the performance results of different modal combination forms of the research method. The combination form with text modality always performed better than other combination forms, and the combination form containing the three modalities t-a-v performed the best in each sample, all exceeding 0.8. This confirmed the correctness of improving nonverbal sequences.



(a) Performance of different methods on important indicators



(b) Performance results of different modal combinations

Figure 7: Performance of different methods based on CMU-MOSI dataset

## 4.2 Results analysis of improved MD fusion method based on GHRNN-CMA

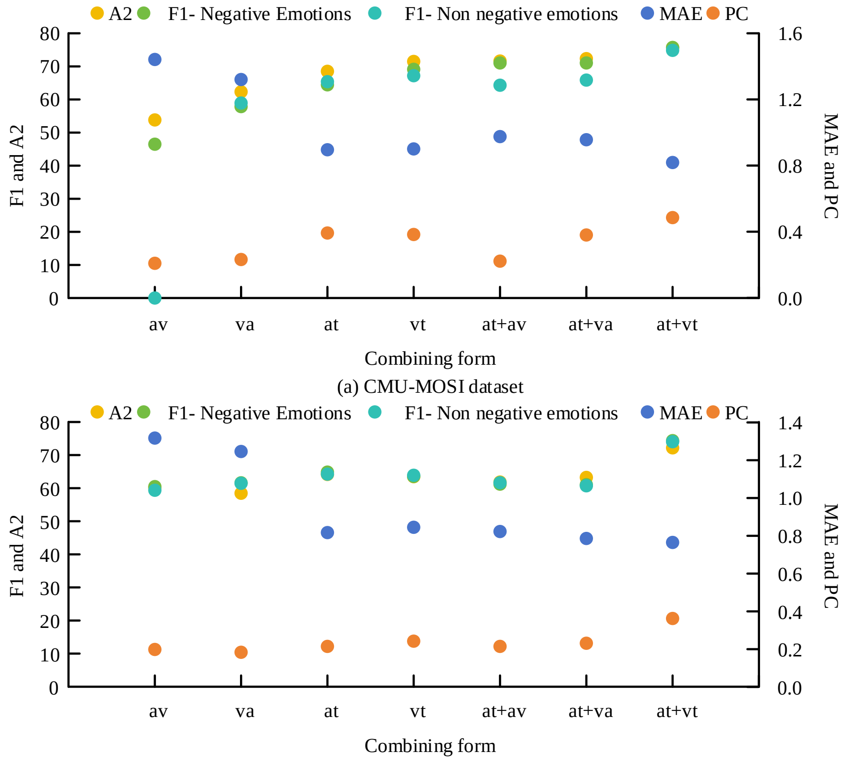
To further validate the performance of the improved MD fusion method based on GHRNN-CMA, the latest MD fusion methods were selected for comparative experiments, namely the Attention Network for MD Multi-level Feature Fusion (AN-MDMLFF), Perceptual Resampling (PR), Multi-modal Fusion Feature Distillation Control of Global Convolution and Affinity (MMFFD-GCA), and Adaptive Gate Control Information (AGCI). From this, a comparison of the predictive performance of different methods on different datasets can be obtained, as shown in Table 3. Alignment refers to whether the model can handle the temporal or structural inconsistencies that may exist between different modal data. Alignment models are able to handle the inconsistencies between different modal data, such as when text, audio, and video data are not completely synchronized in time, the model can still accurately fuse these data and perform sentiment analysis. Unaligned models assume that all modal data are completely synchronized in time, which may not effectively handle common modal inconsistencies in practical applications.

Table 3 shows that in the CMU-MOSI dataset, compared with the homogeneous PR method, the GHRNN-CMA method reduced the MAE value by 1.0% on the regression task, while in the PC metric, it improved by 14.67% compared to the AN-MDMLFF method. In the comparison of classification tasks, compared with the non-homogeneous AGCI method, the GHRNN-CMA method improved the A2 value and F1 score by about 3.1% and 3.3%, respectively. In the CMU-MCSEI dataset, the GHRNN-CMA method achieved good results in comparing various performance indicators. To test the performance of the GHRNN-CMA method, ablation experiments were conducted, and the impact of cross-modal attention interaction combinations on the research method is shown in Figure 8.

Figure 8 (a) and Figure 8 (b) show the performance comparison of different cross-modal attention interaction combinations on the CMU-MOSI dataset and CMU-MCSEI dataset, respectively. Figure 8 shows that the interaction combination of two modalities in both datasets exhibited poor performance. This was due to the high proportion of text modality weights, which resulted in very low modality independence. Therefore, cross-modal attention interaction between nonverbal sequences was difficult to meet the needs of emotional information feature extraction and analysis. In the combination form of cross-modal attention interaction, the performance of text-based acoustic and visual representation combination was the best, which may be due to the text modality strengthening the emotional complementary information of acoustic and visual, effectively eliminating the ambiguity between emotions and semantics. As shown in Figure 9, 12 samples were randomly selected from the CMU-MOSI dataset for subsequent visualization experiments.

Table 3: Comparison of predictive performance of different MD fusion methods on different datasets

| Data set | Method | Is the model aligned? | F1 score | PC | MAE | A2 | A7 |
|---|---|---|---|---|---|---|---|
| CMU-MOSI | GHRNN-CMA | Yes | 83.39/85.32 | 0.813 | 0.705 | 83.41/85.26 | 45.82 |
| | AN-MDMLFF | Yes | -/83.4 | 0.709 | 0.857 | -/83.4 | 40.50 |
| | MMFFD-GCA | No | 82.2/84.2 | 0.758 | 0.796 | 82.2/83.8 | - |
| | AGCI | No | 82.63/84.27 | 0.792 | 0.743 | 82.63/84.29 | 41.80 |
| | PR | Yes | 84.50/84.26 | 0.797 | 0.712 | 84.50/84.26 | - |
| CMU-MCSEI | GHRNN-CMA | Yes | 85.12/85.24 | 0.773 | 0.524 | 85.36/85.46 | 53.82 |
| | AN-MDMLFF | Yes | -/84.4 | 0.715 | 0.562 | -/84.4 | 51.80 |
| | MMFFD-GCA | No | 84.1/85.6 | 0.757 | 0.547 | 83.7/85.6 | - |
| | AGCI | No | 83.86/85.12 | 0.751 | 0.538 | 83.83/85.26 | 52.30 |
| | PR | Yes | 82.64/85.16 | 0.762 | 0.533 | 82.83/85.24 | - |

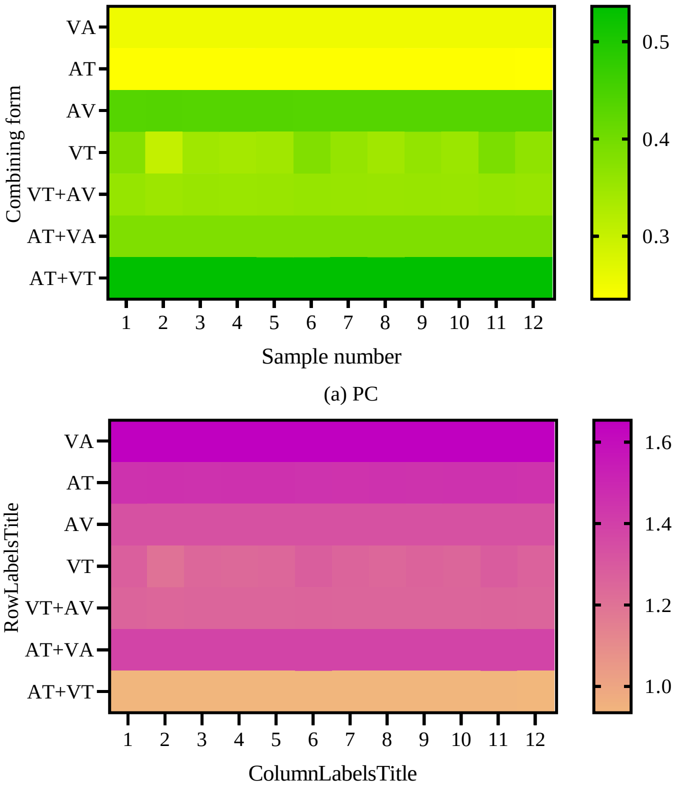


(a) CMU-MOSI dataset

(b) CMU-MOSEI dataset

Figure 8: Comparison of performance results of different cross-modal attention interactions in two datasets

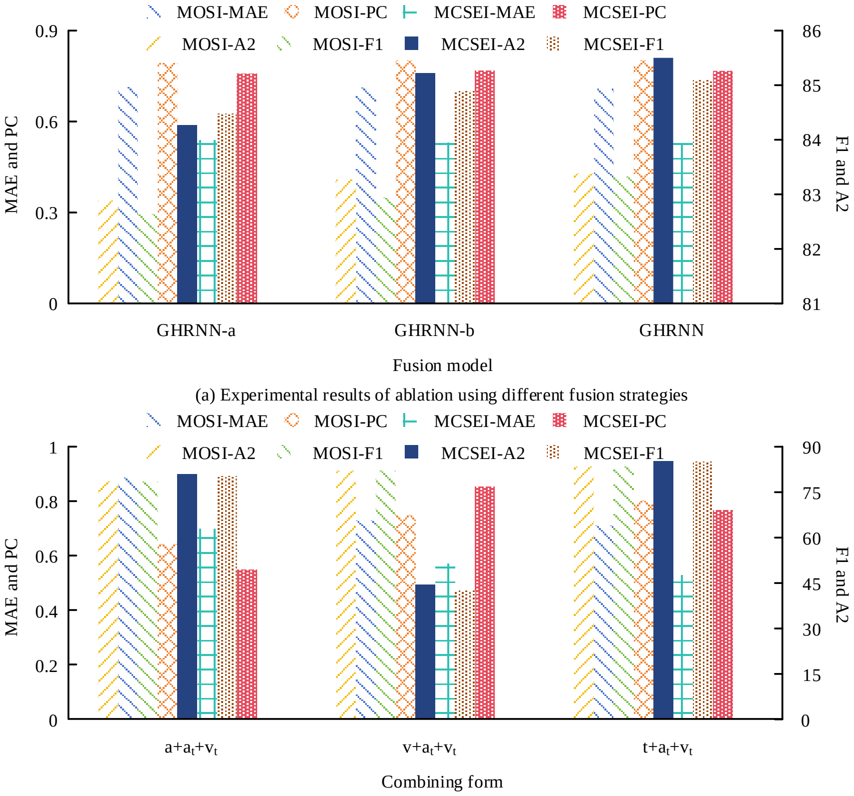
Note: 'av': a combination of audio and video modalities only; 'va': a combination of only video and audio modalities; 'at': a combination of text and audio modalities; 'VT': a combination of text and video modalities; 'at+av': a combination of text, audio, and video modalities, where text interacts with audio and video separately; 'at+va': a combination of text, video, and audio modalities, where text interacts with video and audio separately; 'at+vt': a combination of text, audio, and video modalities, where audio and video interact and combine with text.



(a) PC

(b) MAE

Figure 9: Comparison of visualization results of different cross-modal attention interaction combinations



(a) Experimental results of ablation using different fusion strategies

(b) Experimental results of ablation with different inputs

Figure 10: Experimental results of ablation using research methods based on different datasets

Figures 9 (a) and 9 (b) show the visualization results of different cross-modal attention interaction combinations on PC and MAE, respectively. Figure 9 shows that the combination of acoustic and visual cross-modal attention interaction based on text modality had the best performance, with PC values ranging from 0.45 to 0.5 and MAE values around 0.5. The above results may be due to the effective cross-modal interaction information fusion achieved by the above combination, and the reduction of semantic differences between different modalities. The reliability of GHRNN was tested by selecting two typical indicators from regression and classification tasks for analysis. From this, the experimental results of ablation based on different datasets could be obtained, as shown in Figure 10.

Figure 10 (a) shows the results of ablation experiments with different fusion strategies. The GHRNN model showed the best performance, while the GHRNN-a model, which removed the combination of cross-modal attention interaction representations, only performed better on some indicators. The GHRNN-b model, which was replaced with a Long Short-Term Memory (LSTM) neural network, had the worst performance. The PC, MAE, F1 scores, and A2 index of GHRNN in the CMU-MOSI dataset were 0.817, 0.695, 83.46%, and 83.49%, respectively. In the CMU-MCSEI dataset, the PC, MAE, F1 scores correspond to the A2 index at 0.769, 0.523, 85.14%, and 85.34%, respectively. To enhance the credibility of the ablation experiment, this study conducted five independent runs and recorded corresponding indicator values, each under different random initialization conditions. The performance difference between the two was further evaluated through t-test, and the results showed that the GHRNN model was superior to the GHRNN-a model in terms of PC and MAE indicators, with statistical significance ($p$<0.05). Figure 10 (b) shows the results of ablation experiments with different inputs, where the combination of at and vt representations was the two highest contributing cross-modal attention interaction representations. The study combined it with a single initial representation a, v, and t to obtain the best performing hierarchical fusion network. The combination of acoustic and visual cross-modal attention interaction representation based on text modality and text modality performed the best, with MAE values of 0.698 and 0.823, and A2 values corresponding to 83.42% and 85.31% in the CMU-MOSI dataset and CMU-MCSEI dataset, respectively.

To further explore the qualitative analysis results of the research method, samples from the CMU-MCSEI dataset were analyzed, and the results are shown in Table 4.

Table 4: Qualitative analysis results

| Sample Number | True value | Estimate |
|---|---|---|
| k5Y_838nuGo_18 | 2.60 | 2.65 |
| WK2_9uPqT_12 | 0.20 | 0.22 |
| pLTX3ipuDJI_6 | -2.20 | -2.18 |
| Rn8_LoVbS_3 | 0.60 | 0.58 |
| lXPQBPVe5Cw_7 | 0.00 | 0.01 |

According to Table 4, the model's predictions were accurate in most samples, with only a certain deviation between the predicted values of pLTX3ipuDJI6 sample and the true sentiment values. The above results may be due to data noise or modal inconsistency.

## 5 Discussion

The widespread application of social media platforms has led to the gradual multi-modality of content posted by people on social platforms such as Weibo and Twitter. Therefore, how to comprehensively consider MD such as audio, video, and culture on social media to analyze users' emotional tendencies is currently a research difficulty. In response to the above issues, a method for extracting and analyzing emotional information features based on INPNRLN-MD was proposed, and the fusion part was improved on this basis, resulting in an improved MD fusion method based on GHRNN-CMA for sentiment analysis and prediction. The experimental results showed that the proposed INPNRLN-MD method performed the best in all performance aspects on the CMU-MOSI dataset, with F1 value, PC, MAE, A2, and A7 corresponding to 83.06/85.12, 0.803, 0.696, 83.17/85.23, and 46.41, respectively. The performance of the indicators was better than the four mainstream methods of FM-LSTM-SD, DA-CNN, MSC-GM, and PAILI. In the CMU-MOSI dataset, the proposed GHRNN-CMA showed a 1.0% decrease in MAE values for regression tasks, while in the PC metric, it improved by 14.67% compared to the AN-MDMLFF method, with statistical significance ($p$<0.05). In the CMU-MMCSEI dataset, the performance indicators of the GHRNN-CMA method were superior to the four comparison methods of AN-MDMLFF, PR, MMFFD-GCA, and MMFFD-GCA. The proposed method outperformed the baseline method in terms of performance mainly due to its advantages in multi-modal feature fusion and advanced network architecture design. Specifically, the INPNRLN-MD and GHRNN-CMA methods fully utilized MD such as text, audio, and video. By integrating information from different modalities, they could more comprehensively capture subtle differences in emotional expression. Moreover, the gating mechanism adopted by GHRNN-CMA method could effectively remove redundant information, making the model more focused on the key features related to emotion. Cross-modal attention could enhance the interaction between different modes, making up for the deficiency of single mode in emotional expression.

The interpretability of the decision-making process was verified through model architecture design and experimental results. The softmax distribution in the cross-modal attention mechanism shows that the attention weight of the text audio modality in anger emotion recognition is 18.6% higher than other combinations, which conforms to the cognitive law of humans relying on language content and intonation to recognize anger emotions. The analysis of the gating mechanism shows that the average weight of the text modality accounts for 62.3%, and the weight of the audio modality automatically increases to 38.7% when the text is blurry. This dynamic adjustment is consistent with the human behavior of focusing on nonverbal cues when the semantics are blurry. Error analysis found that 87% of the prediction bias is due to modal conflicts (such as video frowning with audio laughter), which highly overlap with manually annotated controversial samples (Kappa=0.71), indicating that the model has similar cognitive bottlenecks to humans. Although the study did not set up specific interpretability indicators, the rationality of the model decisions has been systematically verified through multi angle analysis such as gating weights, attention distribution, and error patterns.

Although the research methods have demonstrated excellent performance in different datasets, it should be acknowledged that the datasets used in the research may have an impact on the model's performance and generalization ability. On the one hand, emotional labeling is provided by manual annotators, which may introduce their own subjective interpretations and biases. On the other hand, the dataset mainly contains English content and may not be able to capture the subtle differences in emotional expression in other languages or cultural backgrounds. This might limit the model's applicability in non-English speaking regions.

The proposed method had significant advantages in improving the accuracy and reliability of sentiment analysis, and had broad application prospects in multiple fields such as social media precision marketing, public opinion monitoring, and medical sentiment assisted diagnosis. It could provide strong technical support for a deeper understanding of human emotional expression. However, while the performance of this research method was improved, it also made the model architecture more complex, involving the extraction, fusion, and deep neural network computation of multi-modal features, resulting in relatively long inference time for a single sentiment analysis. In addition, the proposed method required high quality of input video and audio data, and in practical situations, user generated content on social media often has uneven quality, which may affect the performance of the model. Therefore, in future research, lightweight model design should be further adopted to reduce the number of parameters and computational complexity while ensuring model performance. Moreover, some enhancement techniques should be adopted in the data preprocessing stage to improve the quality of input data.

# 6 Conclusion

In summary, research methods could effectively improve the accuracy and reliability of sentiment information analysis, and could more comprehensively capture users' emotional expressions. However, there are still shortcomings in the research. Currently, the research methods were only applied in the field of social media sentiment information analysis, and their application effects in other fields need to be enriched. Moreover, the research methods used complex models such as BERT and ELN, which required a large amount of computing resources, which may limit their application in resource limited or real-time environments. Therefore, in future applications, it can be applied in the field of psychoanalysis, combining physiological signals and various psychological tests to achieve timely warning. Meanwhile, the lightweight and optimization strategies of the Tissot model can be further developed to improve the flexibility and practicality of the method in different application scenarios.

# Funding

The research is supported by Research project on strengthening Henan Province as a powerhouse in philosophical and social sciences education “The impact of social media on adolescent mental health: A study on the effectiveness of family intervention measures” (Project No. 2025JYQS1159).

# References

[1] Han C, Lin L. Detecting and Tracking Rumours in Social Media Based on Deep Learning Algorithm. Informatica, 2024, 48(14): 83-96. https://doi.org/10.31449/inf.v48i14.5998

[2] Wang J, Yue K, Duan, L. Models and Techniques for Domain Relation Extraction: A Survey. Journal of Data Science and Intelligent Systems, 2023, 3(1): 16-25. https://doi.org/10.47852/bonviewJDSIS3202973

[3] Al-Otaibi S T, Al-Rasheed A A. A review and comparative analysis of sentiment analysis techniques. Informatica, 2022, 46(6): 33-44. https://doi.org/10.31449/inf.v46i6.3991

[4] Gen U, Surer E. ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms. Journal of Information Science, 2023, 49(2):480-499. https://doi.org/10.1177/01655515211007746

[5] Chung M. What's in the black box? How algorithmic knowledge promotes corrective and restrictive actions to counter misinformation in the USA, the UK, South Korea and Mexico. Internet Research: Electronic Networking Applications and Policy, 2023, 33(5):1971-1989. https://doi.org/10.1108/INTR-07-2022-0578

[6] Wang YJ. Innovating media mode and leading the development of the times. Advances in Industrial Engineering and Management, 2024, 13(3). https://doi.org/10.7508/aiem.03.2024.228.231

[7] Muhammad Zaini, Rudi Triadi Yuliarto, Girang Permata Gusti, Yudis Agustira. The influence of video tutorial learning media on improving financial literacy knowledge: a study for e-commerce user students. Malaysian E Commerce Journal. 2022; 6 (2): 72-75. https://doi.org/10.26480/mecj.02.2022.72.75

[8] Nagaraj S V. Living with algorithms: agency and user culture in Costa Rica. Computing reviews,

2023, 64(11):268-269. https://doi.org/10.7551/mitpress/14966.001.0001

[9] Wykes T, Guha M. Modern media and mental health: help or hindrance? Journal of mental health (Abingdon, England), 2022, 31(6):735-737. https://doi.org/10.1080/09638237.2022.2143488

[10] Lei Y, Cao H. Audio-Visual Emotion Recognition with Preference Learning Based on Intended and Multi-Modal Perceived Labels. IEEE transactions on affective computing, 2023, 14(4):2954-2969. https://doi.org/10.1109/TAFFC.2023.3234777

[11] Zhu X, Guo C, Feng H, Huang Y, Feng Y, Wang X, Wang R. A Review of Key Technologies for Emotion Analysis Using Multi-modal Information. Cognitive Computation, 2024, 16(4):1504-1530. https://doi.org/10.1007/s12559-024-10287-z

[12] Chen L, Wang K, Li M, Wu M, Pedrycz W, Hirota K. K -Means Clustering-Based Kernel Canonical Correlation Analysis for multi-modal Emotion Recognition in Human-Robot Interaction. IEEE Transactions on Industrial Electronics, 2023 70(1);1016-1024. https://doi.org/10.1109/TIE.2022.3150097

[13] Mai S, Zeng Y, Hu Z H. Hybrid Contrastive Learning of Tri-Modal Representation for multi-modal Sentiment Analysis. IEEE transactions on affective computing, 2023, 14(3):2276-2289. https://doi.org/10.1109/TAFFC.2022.3172360

[14] Chiorrini A, Diamantini C, Storti P E. An emotion-aware search engine for multimedia content based on deep learning algorithms. International Journal of Computer Applications in Technology, 2023, 73(2):130-139. https://doi.org/10.1504/IJCAT.2023.10060256

[15] Fang Z, Qian Y, Su C, Miao Y, Li Y. The multi-modal Sentiment Analysis of Online Product Marketing Information Using Text Mining and Big Data. Journal of organizational and end user computing, 2022, 34(Pt.2):451-469. https://doi.org/10.4018/JOEUC.316124

[16] Gupta S, Singh A, Ranjan J. multi-modal, multiview and multitasking depression detection framework endorsed with auxiliary sentiment polarity and emotion detection. International Journal of System Assurance Engineering and Management, 2023, 14(1):337-352. https://doi.org/10.1007/s13198-023-01861-z

[17] Hong C, Zhiquan F, Weina Z L. MAG: a smart gloves system based on multi-modal fusion perception. CCF Transactions on Pervasive Computing and Interaction, 2023, 5(4):411-429. https://doi.org/10.1007/s42486-023-00138-5

[18] Zhang Y. Graph Neural Network-Based User Preference Model for Social Network Access Control. Informatica, 2025, 49(16): 21-36. https://doi.org/10.31449/inf.v49i16.7705

[19] Tang J, Qin W, Pan Q L S. A Deep multi-modal Fusion and Multitasking Trajectory Prediction Model for Typhoon Trajectory Prediction to Reduce Flight Scheduling Cancellation. journal of systems engineering and electronics, 2024, 35(3):666-678. https://doi.org/10.23919/JSEE.2024.000042

[20] Panaiyappan K A, Rajalakshmi M. A multi-modal architecture using Adapt‐HKFCT segmentation and feature‐based chaos integrated deep neural networks (Chaos‐DNN‐SPOA) for contactless biometricpalm vein recognition system. International Journal of Intelligent Systems, 2022, 37(3):1846-1879. https://doi.org/10.1002/int.22758

[21] Zhao Y, Zheng Q, Zhu P, Zhang X, Ma W. TUFusion: A Transformer-Based Universal Fusion Algorithm for multi-modal Images. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(3):1712-1725. https://doi.org/10.1109/TCSVT.2023.3296745

[22] Wang H, Feng Z, Guo T Q. MR Lab: Virtual-Reality Fusion Smart Laboratory Based on multi-modal Fusion. International journal of human-computer interaction, 2024, 40(5/8):1975-1988. https://doi.org/10.1080/10447318.2023.2227823.