ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20230628001154
Original scientific paper

Emotion Intensity Detection in Online Media: An Attention Mechanism Based Multimodal Deep Learning Approach

Yuanchen CHAI

Abstract: With the increasing influence of online public opinion, mining opinions and trend analysis from massive data of online media is important for understanding user sentiment, managing brand reputation, analyzing public opinion and optimizing marketing strategies. By combining data from multiple perceptual modalities, more comprehensive and accurate sentiment analysis results can be obtained. However, using multimodal data for sentiment analysis may face challenges such as data fusion, modal imbalance and inter-modal correlation. To overcome these challenges, the paper introduces an attention mechanism to multimodal sentiment analysis by constructing text, image, and audio feature extractors and using a custom cross-modal attention layer to compute the attention weights between different modalities, and finally fusing the attention-weighted features for sentiment classification. Through the cross-modal attention mechanism, the model can automatically learn the correlation between different modalities, dynamically adjust the modal weights, and selectively fuse features from different modalities, thus improving the accuracy and expressiveness of sentiment analysis.

Keywords: attention mechanism; emotion detection; multimodal; online media

1 INTRODUCTION

Nowadays, the influence of online public opinion is getting bigger and bigger, and more and more netizens can express their emotional color and emotional tendency on social media, including their opinions on commodities, characters, social events, etc. [1, 2]. Therefore, sentiment analysis of text and mastering the trend of online public opinion have become important elements of natural language processing. Therefore, how to mine opinions and tendency analysis from the massive data of online media is of great significance for several aspects [3]. 1) Online media sentiment analysis can help us understand users' emotional tendencies in online media such as social media, news websites, and comment platforms. By analyzing user sentiment, we can better understand their attitudes and emotional responses to specific topics, products, events, or opinions. 2) Online media sentiment analysis can help companies and brands monitor and manage their online reputation. By analyzing user sentiment on social media and review platforms, companies can identify and respond to negative sentiments in a timely manner to protect their brand reputation. 3) Online media sentiment analysis can help governments, organizations, and public institutions understand the public's emotional attitudes toward specific events, policies, or social issues. By analyzing sentiment on social media and news sites, public opinions and emotional tendencies can be better understood. 4) Online media sentiment analysis can help marketers and advertisers understand users' emotional responses to advertisements, marketing campaigns, and product promotions. By analyzing user sentiment, advertising and marketing strategies can be optimized to improve advertising effectiveness and user engagement [4, 5].

In summary, sentiment analysis of online media is important for understanding user sentiment, managing brand reputation, analyzing public opinion and optimizing marketing strategies. By accurately identifying and understanding user sentiment, we can better meet user needs and provide better products and services, while improving brand image and user experience. However, nowadays simple text sentiment detection algorithms can no longer meet the needs of complex Internet systems, and multimodal fusion sentiment classification methods have gradually become mainstream [6, 7].

This is because multimodal data combines different perceptual modalities such as text, image, audio and video. Each perceptual modality can provide unique sentiment information. By combining data from multiple perceptual modalities, more comprehensive and accurate sentiment analysis results can be obtained. For example, in sentiment analysis, images can provide nonverbal features such as facial expressions and body language, and audio can provide sound features such as tone and volume, and this information can enrich the sentiment analysis. Multimodal data can also provide more contextual information to help us better understand and interpret sentiment. For example, in sentiment analysis on social media, relying on text alone may not be able to fully capture the true emotions of users. However, combining multimodal data, such as images and audio, can better understand the user's emotional expression and thus improve the accuracy of sentiment analysis.

However, using multimodal data for sentiment analysis may face the following challenges: 1) Data fusion: multimodal data usually comes from different perceptual modalities, such as text, image, audio, etc. Effective fusion of these different data is a challenge. Suitable methods need to be designed to integrate data from different modalities to better capture emotional information. 2) Modal imbalance: there may be an imbalance in the data from different perceptual modalities, i.e., certain modalities have less data. This may lead to insufficient data learning for certain modalities in the training process of the model, affecting the accuracy of sentiment analysis. Appropriate strategies are needed to deal with modal imbalance, such as data augmentation, sample weighting, etc. 3) Inter-modal correlation: data from different perceptual modalities may have certain correlations, e.g., the sentiment in an image may be correlated with the sentiment in a text. How to effectively utilize the correlation between different modalities is a challenge. Appropriate model structures and algorithms need to be

designed to capture and utilize inter-modal correlation information [8, 9].

To overcome the above challenges, we introduce the attention mechanism into multimodal sentiment analysis, specifically, in this paper, we first construct a text feature extractor, an image feature extractor, and an audio feature extractor respectively. Then, we use a custom cross-modal attention layer to compute the attention weights between different modalities. Finally, we fuse the attention-weighted features and add an output layer for sentiment classification [10].

The main contributions of this paper include the following: Unlike traditional attention mechanisms that typically focus on single-modality data, our algorithm is uniquely designed to handle the complexities and nuances of multimodal data text, images, and audio. It dynamically computes attention weights between different modalities, enabling the model to adaptively focus on the most salient features across these diverse data types. This adaptability is crucial in accurately detecting emotion intensity, as it allows the model to seamlessly integrate and prioritize information from different sources based on their contextual relevance. Furthermore, this cross-modal attention mechanism offers a more efficient way of processing multimodal data by mitigating the challenges of data redundancy and irrelevance that often plague multimodal systems. By selectively focusing on pertinent features and effectively fusing them, our model achieves higher accuracy and efficiency in emotion detection. This approach not only enhances the model's performance in accurately classifying emotions but also significantly reduces computational overhead, making it more suitable for real-time applications. The contrast with existing models lies in this unique capacity to intelligently navigate and synthesize complex multimodal data landscapes, thereby pushing the boundaries of emotion detection technology [11, 12].

2 RELATED WORKS

2.1 Sentiment Analysis of Multimodal Data

There are several main methods for sentiment analysis of multimodal data [13]: 1) Fusion feature methods: the features of different modalities are fused, and then sentiment analysis is performed using traditional machine learning algorithms or deep learning models. Common fusion feature methods include feature-level fusion and decision-level fusion [13]. 2) Multimodal feature learning methods: mapping features of different modalities into a shared feature space by learning the correlation between multimodal data. Common multimodal feature learning methods include Principal Component Analysis (PCA), Autoencoder, and Multi-view Learning, etc. 3) Cross-modal Attention Mechanism Methods: Attention mechanisms are utilized to automatically learn the degree of attention between different modalities, so as to perform weighted fusion of features from different modalities. Cross-modal attention mechanism methods can be realized by dot product attention, bilinear attention, or multi-head attention, etc. [14]. 4) Cross-modal Pre-training Methods: By pre-training on large-scale multimodal data, cross-modal representation capabilities are learned and then fine-tuned on specific tasks. Common cross-modal pre-training methods include Multimodal Autoencoder and cross-modal pre-training models (e.g., BERT, GPT, etc.) [15]. 5) Graph Neural Network Methods: Graph Neural Networks are utilized to model the relationships between multimodal data for sentiment analysis. Graph neural networks can effectively capture the structural and semantic information of multimodal data [16].

Multimodal data usually contains different types of information, such as text, images, audio, etc. The attention mechanism can help the model automatically learn the level of attention between different modalities and determine how much each modality contributes to the sentiment analysis [13]. This can better utilize the information of different modalities and improve the accuracy and performance of sentiment analysis. In addition, multimodal data sentiment analysis requires the fusion of features from different modalities to synthesize information from multiple modalities [15]. The attention mechanism can weight the fusion of features from different modalities based on the correlation between modalities. This can better capture the relationship and importance between different modalities and improve the expressive power of sentiment analysis [15]. The importance of different modalities in sentiment analysis may vary with the task and data. The attention mechanism can dynamically adjust the weights of modalities according to specific tasks and data, so that the model can adaptively focus on the information of different modalities. This can improve the flexibility and adaptability of the model. Therefore, choosing the attention mechanism for multimodal data sentiment analysis can help the model better utilize the information of different modalities and improve the accuracy, expressiveness and flexibility of sentiment analysis. The attention mechanism can also provide interpretability and explainability of model decisions, increasing understanding and trust in the model [14, 16].

2.2 Multimodal Attention Mechanisms

Multimodal Attention Mechanism is a method for sentiment analysis of multimodal data, which helps the model to automatically learn the level of attention between different modalities and weight the fusion of features of different modalities. There are several main multimodal attention mechanism methods: 1) Dot Product Multimodal Attention (Dot Product Based Multimodal Attention): measures the correlation between different modalities by calculating the dot product between them, and weights the correlation as a weight to weighted fusion of features from different modalities. This method can capture the interactions between different modalities [17]. 2) Bilinear Multimodal Attention: calculates the correlation between different modalities by using a bilinear function, and weights the correlation as a weight to weight the fusion of features from different modalities. Bilinear Multimodal Attention can capture more complex modal relationships [18]. 3) Cross-Modal Adaptive Attention: By learning the parameters of the attention weights, the model can adaptively adjust the degree of attention between different modalities. This approach can dynamically adjust the weights of modalities according to specific tasks and data [19]. 4) Cross-Modal Multi-Head Attention: Apply the

attention mechanism to multiple Heads, each of which can learn a different degree of modal attention. Cross-Modal Multi-Head Attention can improve the modelling ability of the model on multimodal data [20]. 5) Multimodal Attention in Graph Neural Networks: In Graph Neural Networks, by introducing the Multimodal Attention mechanism, the nodes of different modalities can be weighted and fused to perform the Multimodal data sentiment analysis [21].

Sentiment analysis of multimodal data involves interactions and associations between different modalities. The cross-modal attention mechanism can help the model automatically learn the degree of association between different modalities and weightedly fuse the features of different modalities according to the degree of association [21]. This can better capture the interaction information between different modalities and improve the accuracy and expressiveness of sentiment analysis. Second, multimodal data sentiment analysis requires fusion of features from different modalities to comprehensively consider information from multiple modalities. The cross-modal attention mechanism can weight the fusion of features from different modalities according to the correlation between modalities. This can better capture the relationship and importance between different modalities and improve the performance and accuracy of sentiment analysis. In addition, the importance of different modalities in sentiment analysis may change with the task and data. The cross-modal attention mechanism can dynamically adjust the weights of modalities according to specific tasks and data, so that the model can adaptively focus on the information of different modalities. This can improve the flexibility and adaptability of the model. Therefore, choosing the cross-modal attention mechanism for multimodal data sentiment analysis can help the model better utilize the information of different modalities and improve the accuracy, expressiveness and flexibility of sentiment analysis [17].

3 METHODS

3.1 Overall Framework

Each modality serves a unique function in capturing and interpreting emotional cues, contributing to a holistic understanding of emotional states. Textual data analysis involves employing Natural Language Processing (NLP) techniques to discern emotional undertones from linguistic patterns, such as word choice, syntax, and semantic context. This approach effectively captures explicit emotional expressions and subtle linguistic cues indicative of deeper emotional states. Image analysis, on the other hand, focuses on visual cues such as facial expressions, body language, color usage, and object placements within images. By leveraging advanced image processing techniques, the model interprets these visual elements to infer emotional contexts, providing insights into unspoken emotional cues. Audio data analysis complements these modalities by examining vocal features, including tone, pitch, volume, and speech rate. Through sophisticated audio signal processing, the model decodes auditory cues that reflect emotional states, often conveying nuances that text or images alone might miss.

The integration of these modalities is achieved through a cross-modal attention mechanism, which dynamically evaluates and weights the input from each modality. This process not only ensures that the most relevant features from each data type are considered but also enables the model to adapt to the varying significance of each modality across different contexts. For instance, in scenarios where textual data is sparse or ambiguous, the model can lean more heavily on audio and visual cues, and vice versa. This integrated approach allows for a more nuanced and comprehensive analysis of emotion intensity, surpassing the capabilities of single-modality systems. By leveraging the strengths of each modality and intelligently combining them, our model provides a robust and versatile solution for emotion intensity detection in diverse online media environments. This holistic approach is particularly effective in complex scenarios where emotions are conveyed through a rich tapestry of textual, visual, and auditory signals, ensuring a more accurate and profound understanding of emotional states.

The overall framework of the algorithm developed in this paper is shown in Fig. 1. The framework is a system for cross-modal emotion detection, including three feature extractors, text, speech and video, and a cross-modal attention mechanism. Among them, the text feature extractor uses a bidirectional LSTM to capture the contextual information of the text, the speech feature extractor uses a convolutional neural network (CNN) to extract the local features of the speech signal, and the video feature extractor uses a three-dimensional convolutional neural network (3D CNN) to simultaneously consider the spatio-temporal information of the video. The cross-modal attention mechanism fuses text, speech, and video features by calculating the attention weights to obtain the final cross-modal feature representation. Finally, multiclassification is performed by multilayer perceptron and softmax function and sentiment classification is performed using multiclassification cross-entropy loss function.

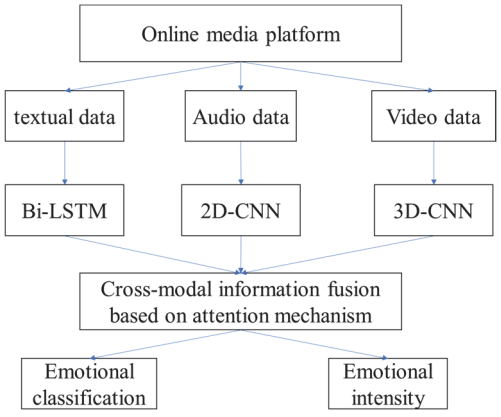

Figure 1 Overall framework

3.2 Text Feature Extractor

Bi *LSTM* is a variant of Recurrent Neural Network (*RNN*) for text feature extraction. It captures contextual information in text by running two *LSTM* (Long Short-Term Memory) layers in both forward and reverse directions in time.

The principle of Bi - *LSTM* is as follows: first, the text sequence is converted into word embedding vector representations, where each word embedding vector represents a word feature. Second, the sequence of word embedding vectors is input into a forward *LSTM*, which processes the input sequence in sequence and outputs a hidden state at each time step. After that, the sequence of word embedding vectors is reversed and input into the reverse *LSTM*, which processes the input sequence in reverse order and outputs one hidden state per time step. Finally, the hidden states of forward *LSTM* and reverse *LSTM* are spliced to obtain a comprehensive feature representation for each time step.

The formula of Bi - *LSTM* is as follows: assume the input sequence is $X = [x_1, x_2, ..., x_n]$, where x_i denotes the word embedding vector of the ith word. the forward propagation process of BiLSTM can be expressed as:

$$h_i^f = LSTM_f\left(x_i, h_{i-1}^f, c_{i-1}^f\right) \tag{1}$$

Where $LSTM_f$ denotes the update rule of the forward *LSTM*, and h_{i-1}^f and c_{i-1}^f denote the hidden state and cell state of the previous time step of the forward *LSTM*, respectively. The propagation process of the reverse *LSTM* can be expressed as:

$$h_i^b = LSTM_b\left(x_i, h_{i+1}^b, c_{i+1}^b\right) \tag{2}$$

where $LSTM_b$ denotes the update rule of the reverse *LSTM*, and h_{i-1}^b and c_{i-1}^b denote the hidden and cellular states of the latter time step of the reverse *LSTM*, respectively. The hidden states of forward *LSTM* and reverse *LSTM* are spliced to obtain the integrated feature representation:

$$h_i = \left(h_i^f, h_i^b\right) \tag{3}$$

Finally, the integrated feature representation can be input into the multimodal fusion attention mechanism. Fig. 2 shows the Bi - *LSTM* network.

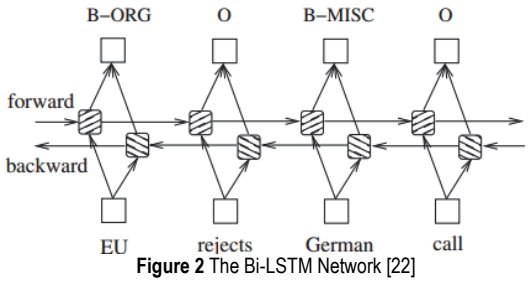

Figure 2 The Bi-LSTM Network [22]

3.3 Voice Feature Extractor

Since CNNs perform well in processing image data and speech signals can be regarded as a two-dimensional data where one dimension is time and the other is frequency. Therefore, in this paper, CNN is used as a speech feature extractor, CNN can effectively extract local features in speech signals through the combination of convolutional and pooling layers. This is important for localized information such as spectral shape and frequency variation in speech signals. In addition, CNN has the property of translational invariance, which means it can detect the same features in different time periods. For speech signals, this is very useful because the features of speech signals are usually independent of their specific position on the time axis. The parameter sharing mechanism in CNNs also reduces the number of parameters in the model and improves the efficiency of model training. Speech signals usually have multiple frequency channels, each corresponding to a different frequency range. CNNs can process the features of multiple channels simultaneously with multiple convolutional kernels, thus capturing the spectral information of the speech signal more comprehensively.

Suppose we have an input $x(t)$ of a speech signal, where t denotes time. We wish to extract its features by CNN. The input signal is first sub framed to divide it into multiple time windows, and a Fourier transform is applied to each time window to convert the signal into a spectral representation. Then the signal is fed to the convolutional layer whose input is a spectral representation of the speech signal of size (N, C, F, T), where N is the batch size, C is the number of channels, F is the frequency dimension, and T is the time dimension. The convolution kernel is parameterized by W with size (C_{out}, C, K_f, K_t), where C_{out} is the number of output channels, K_f is the size of the convolution kernel in the frequency dimension, and K_t is the size of the convolution kernel in the time dimension. The bias is b with size (C_{out}). The output is Y with size $(N, C_{out}, F_{out}, T_{out})$, where F_{out} is the output frequency dimension and T_{out} is the output time dimension. Namely,

$$Y_{n,c_f,t} = \sum_{c_i n=1}^{C} \sum_{k_f=1}^{K_f} \sum_{k_t=1}^{K_t} X_{n,c_i n,f+k_f,t+k_t} \times \times W_{c,c_i n,k_f,k_t} + b_c \tag{4}$$

Where X is the input, W is the convolution kernel, and b is the bias.The output of the convolutional layer is then fed to the pooling layer, the output of the convolutional layer, Y, of size $(N, C_{out}, F_{out}, T_{out})$. The pooled feature representation Z, of size $(N, C_{out}, F_{pool}, T_{pool})$, where F_{pool} is the frequency dimension after pooling and T_{pool} is the time dimension after pooling. Namely.

$$Z_{n,c,f,t} = Pool\left(Y_{n,c,f,t}\right) \tag{5}$$

where Pool denotes the pooling operation.

3.4 Video Feature Extractor

Video feature extractor is 3D Convolutional Neural Network (3D CNN) and the main reason for using it for video feature extraction is the ability to consider both temporal and spatial information. Video data is a temporal data that contains consecutive frames and each frame contains spatial information. Traditional 2D Convolutional

Neural Networks (2D CNNs) can only handle static images and cannot capture the temporal dependencies and dynamic features in videos. By introducing convolutional operations in the temporal dimension, 3D CNN can perform convolutional operations in time and space simultaneously, thus being able to capture the dynamic features in the video. 3D CNN can perform convolutional operations on each frame in the video sequence and perform sliding window operations in the temporal dimension, thus extracting the spatio-temporal features in the video. Using 3D CNN for video feature extraction can take into account the temporal dependencies in the video, capture the motion information in the video and reduce the number of parameters to improve the efficiency and generalization of the model.

Suppose we have a video sequence containing T frames, each with dimensions $H \times W$ (height × width) and the number of channels C. We can represent the video sequence as a four-dimensional tensor X of shape (T, C, H, W). First the video sequence X is taken as input. Then a convolution operation is performed on the video sequence using a 3D convolution kernel to capture temporal and spatial features. Suppose we use K 3D convolution kernels, each with dimensions $D \times D \times D$ where D denotes the depth, height and width of the convolution kernel. For each convolutional kernel, we can obtain a feature map with the shape (K, H', W'), where H' and W' are the height and width of the feature map after convolution. In addition, the output of the convolutional layer needs to be nonlinearly transformed, and common activation functions include ReLU, Sigmoid and Tanh. Then, downsampling of the spatial dimension of the feature map is performed to reduce the number of parameters and extract more robust features. Common pooling operations include maximum pooling and average pooling. Finally, the output of the pooling layer is spread into a one-dimensional vector, and feature mapping and classification are performed through a fully connected layer. The fully connected layer can include multiple hidden layers and output layers, and finally the feature representation of the video is output. Fig. 3 shows the 3D convolution operation.

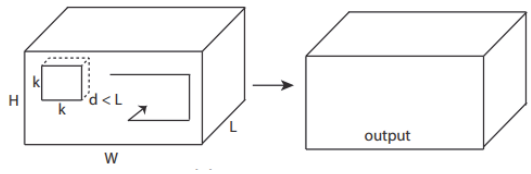

Figure 3 3D convolution operation [23]

3.5 Cross-modal Attention Mechanism and Emotion Detection

Suppose we have a text feature representation as T, a speech feature representation as A, and a video feature representation as V. We would like to fuse these features through an attention mechanism to get the final cross-modal feature representation F. We can use a fully connected layer and an activation function (e.g., ReLU) to compute the text weights.

1) Calculate the attention weights: for text feature T, we can use a fully connected layer and an activation function (e.g., ReLU) to calculate the attention weights $T_{att} = f(T)$ for the text. For speech feature A, we can use a fully connected layer and an activation function to compute the attentional weight $A_{att} = f(A)$ for speech. For video feature V, we can use a fully connected layer and an activation function to compute the attentional weight for video $V_{att} = f(V)$.

2) Normalized Attention Weights: For the attention weight T_{att} of text, we can use the softmax function to normalize it and get the normalized attention weight:

$$T_{att_{norm}} = soft\max(T_{att}) \tag{6}$$

For the attention weight A_{att} for speech, we can use the softmax function for normalization to get the normalized attention weight:

$$A_{att_{norm}} = soft\max(A_{att}) \tag{7}$$

For the video attention weight V_{att}, we can use the softmax function for normalization to get the normalized attention weight:

$$V_{att_{norm}} = soft\max(V_{att}) \tag{8}$$

3) Weighted fusion feature: multiply text feature T and normalized text attention weight $T_{att_{norm}}$ by elements to get weighted text feature $T_{fused} = T \times T_{att_{norm}}$. Multiply speech feature A and normalized speech attention weight $A_{att_{norm}}$ by elements to get weighted speech feature $A_{fused} = A \times A_{att_{norm}}$. Multiply the video feature V with the normalized video attention weight $V_{att_{norm}}$ on an element-wise basis to obtain the weighted video feature $V_{fused} = V \times V_{att_{norm}}$.

4) Final cross-modal feature representation: the weighted text feature T_{fused}, speech feature A_{fused} and video feature $V_{att_{norm}}$ are summed by elements to get the final cross-modal feature representation:

$$F = T_{fused} + A_{fused} + V_{fused} \tag{9}$$

After that, F is fed into a multilayer perceptron, and then a softmax function is utilized to compute the probability of multiclassification, and the loss function for sentiment classification is the multiclassification cross-entropy loss function.

4 EXPERIMENTAL RESULTS

4.1 Data

The dataset used in this paper is a multimodal sentiment analysis dataset. The dataset aims to study the

expression and recognition of feelings, opinions and emotions in multimodal data. The dataset contains movie clips, speech and text data from YouTube. These data clips cover a wide range of emotion and opinion expressions, including speech and dialog. Each sample in the dataset contains data in multiple modalities, including video, speech, and text.

This dataset is particularly chosen for its comprehensive coverage of emotion and opinion expressions across multiple modalities video, speech, and text predominantly sourced from YouTube. This variety enables an in-depth analysis of emotions from a multimodal perspective, crucial for our study's aim to develop a cross-modal emotion detection framework.

Each sample in the dataset is meticulously labeled for emotional categories such as joy, sadness, anger, etc., and, importantly, also for the intensity of these emotions. These dual labels are vital for our research, as they allow not only for categorizing emotions but also for assessing their intensity, providing a nuanced understanding of emotional expressions. This is particularly relevant for our model, which aims to discern not just the type but also the depth of emotions from multimodal data.

Furthermore, the large-scale nature of the dataset, with its extensive collection of samples, ensures a robust and comprehensive analysis. It allows our model to be trained and tested across a diverse range of emotional expressions, ensuring that it is well-equipped to handle real-world complexities and variations in emotion detection. This aspect is critical for the generalizability and applicability of our findings.

4.2 Experimental Implementation

The experimental environment for this paper is Tensorflow 2.0, python 3, and the hardware is an RTX2080 GPU with a quad-core intel i7-7700 processor. Due to GPU memory limitation, the batch size is set to 32, and the initial learning rate is 0.001 with adaptive fading processed by Adam optimizer.

Assessment metrics for sentiment categorization include Accuracy: Accuracy measures the proportion of sentiment categories that are correctly predicted by the model. It is one of the most commonly used metrics and is particularly useful for balancing the distribution of categories. Precision: Precision measures how accurately the model predicts a sentiment category. It calculates the proportion of samples predicted to belong to a sentiment category that actually belongs to that category. Recall: Recall measures the model's ability to recognize an emotion category. It calculates the proportion of samples that actually belong to an emotion category that are correctly predicted by the model to belong to that category. *F*1 Score: *F*1 Score is the harmonic mean of Precision and Recall, which is a combination of the model's accuracy and recognition ability. It is a comprehensive assessment metric for the case of unbalanced category distribution.

The assessment metrics for emotion intensity detection include Mean Squared Error (*MSE*): the *MSE* measures the average squared difference between predicted and actual emotion intensity. A lower *MSE* value indicates that the model's prediction is closer to the actual emotional intensity. Root Mean Squared Error (*RMSE*): the *RMSE* is the square root of the *MSE*, which provides a measure that matches the units of the actual sentiment intensity. Similar to *MSE*, lower *RMSE* values indicate that the model's predictions are closer to actual sentiment intensity. Mean Absolute Error (*MAE*): *MAE* measures the average absolute difference between predicted and actual sentiment intensity. A lower *MAE* value indicates that the model's predictions are closer to the actual sentiment intensity [24-26].

4.3 Comparisons with Baseline Model

The baseline models for multimodal emotion detection in text, speech and video for detecting performance contrasts include: 1) Early Fusion Model: This model fuses features from text, speech and video and then inputs them into a unified classifier for emotion classification. Neural networks (e.g., multilayer perceptron) or other classification algorithms (e.g., support vector machines) can be used as classifiers. Fusion can be achieved by simple splicing, summing or averaging. 2) Late Fusion Model: This model classifies the emotions of text, speech and video separately and then fuses the classification results of each modality. Fusion can be done using methods such as voting mechanism, weighted averaging or decision cascading. For example, the classification results of each modality can be voted and the emotion category with the most votes can be selected as the final prediction. 3) Multi-Modal Deep Learning Model with Parallel Processing: This model uses deep learning methods to input features from text, speech and video into separate neural networks for emotion classification and then fuses the classification results from each modality. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) or other deep learning structures can be used to process multimodal data. The fusion can be achieved by a simple weighted average or voting mechanism.

Table 1 Comparison with baseline model (classification tasks)

Method	Precise	Recall	*F*1	Accuracy
Early Fusion	73.89%	79.13%	74.77%	76.84%
Late Fusion	77.69%	76.35%	77.93%	77.28%
Parallel Processing	79.20%	75.04%	77.32%	76.22%
Proposed method	79.62%	81.47%	79.65%	82.84%

Tab. 1 shows the comparison with baseline model in classification tasks. According to Tab. 1 it can be seen that: Early Fusion method has a precision of 73.89%, a recall of 79.13%, an *F*1 value of 74.77% and an accuracy of 76.84%. Late Fusion method has a precision of 77.69%, recall of 76.35%, *F*1 value of 77.93% and accuracy of 77.28%. Parallel Processing method has a precision of 79.20%, recall of 75.04%, *F*1 value of 77.32% and accuracy of 76.22%. Proposed method has a precision of 79.62%, recall of 81.47%, *F*1 value of 79.65% and accuracy of 82.84%. Based on these evaluation metrics, it can be seen that Proposed method shows better results in terms of precision, recall, *F*1 value and accuracy, and has higher performance compared to the other three baseline models.

Table 2 Comparison with baseline model (regression tasks)

Method	*MSE*	*RMSE*	*MAE*
Early Fusion	0.596	0.801	0.595
Late Fusion	0.615	0.833	0.583
Parallel Processing	0.545	0.758	0.550
Proposed method	0.515	0.751	0.522

Tab. 2 shows the comparison with baseline model in regression tasks. According to Tab. 2, Early Fusion method has *MSE* of 0.596, *RMSE* of 0.801 and *MAE* of 0.595. Late Fusion method has *MSE* of 0.615, *RMSE* of 0.833 and *MAE* of 0.583. Parallel Processing method has *MSE* of 0.545, *RMSE* of 0.758 and *MAE* of 0.550. *RMSE* is 0.758 and *MAE* is 0.550. Proposed method has *MSE* of 0.515, *RMSE* of 0.751 and *MAE* of 0.522. Based on these evaluation metrics, it can be seen that the Proposed method exhibits better results in terms of *MSE*, *RMSE* and *MAE* with respect to the other three methods with lower error values. This suggests that Proposed method may have better prediction performance in this task.

4.4 Parametric Analysis

Fig. 4 shows the matrix of the accuracy of different fusion methods in different modalities (text, audio, video). The number in each cell indicates the accuracy of the corresponding fusion method and modality. According to the data, it can be seen that the "Text + Audio + Video" fusion method has the highest accuracy rate of 82.84% in all modalities. Therefore, "Text+Audio+Video" is considered the best fusion method.

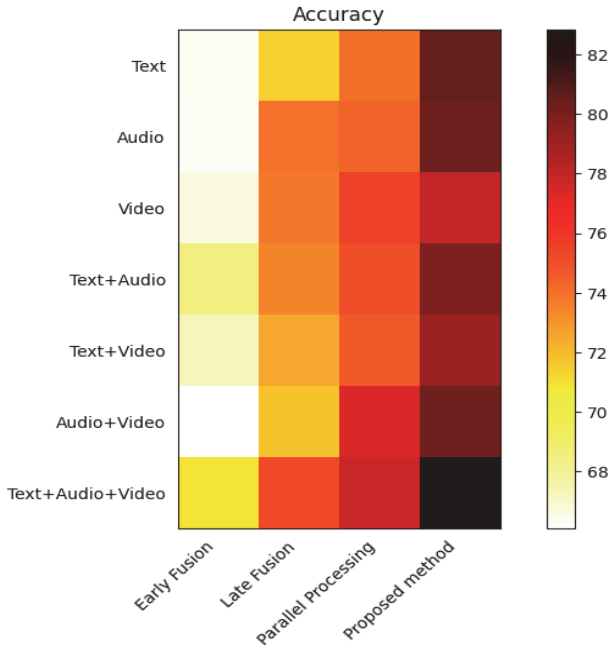

Figure 4 Results for different multimodalities (accuracy)

Fig. 5 shows the *MSE* (Mean Square Error) matrix for different fusion methods in different modalities (text, audio, video). The number in each cell indicates the *MSE* value of the corresponding fusion method and modality. It can be seen that the "Text + Audio + Video" fusion method has the lowest *MSE* value, i.e., 0.515 in different modalities, and therefore, "Text + Audio + Video" is considered to be the most effective fusion method. In conclusion, the fusion of information from the three modal data can be optimized with the method proposed in this paper.

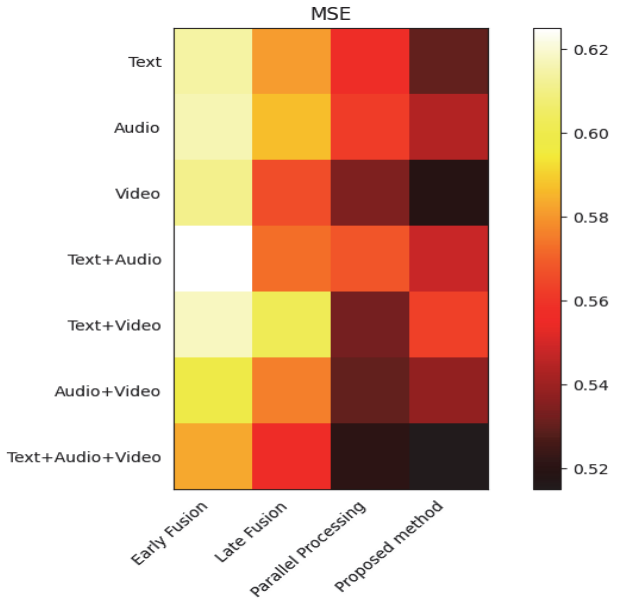

Figure 5 Results for different multimodalities (*MSE*)

5 DISSCUSIONS

5.1 Implications

Using cross-modal attention mechanisms to process text, sound, and video data to accomplish sentiment detection has the following potential implications for human society [24]: 1) Sentiment analysis applications: sentiment detection can help us understand and analyze human emotional states. Processing multimodal data through the cross-modal attention mechanism can capture and understand human emotions more comprehensively, thus providing more accurate and detailed results for sentiment analysis applications. 2) Social media analytics: text, sound, and video data on social media platforms contain a large amount of sentiment information. Processing these data using cross-modal attention mechanisms can help us better understand users' emotional expressions on social media, thus providing more intelligent social media analysis and personalized recommendation services. 3) Emotion recognition and assistance: emotion detection can be applied in the field of emotion recognition and assistance. For example, in the medical field, by analyzing patients' voice, video, and text data, it can help doctors better understand patients' emotional states, and thus provide more accurate diagnosis and treatment plans. 4) Emotion-driven intelligent systems: by using the cross-modal attention mechanism to process multimodal data, it can endow intelligent systems with emotion-aware and emotion-responsive capabilities. Such an intelligent system can better understand and adapt to human emotional needs and provide a more personalized and emotional user experience [25, 27, 28].

5.2 Limitations

Using cross-modal attention mechanisms to process text, sound, and video data to accomplish sentiment

detection may face the following more detailed limitations: 1) Data diversity and generalization ability: cross-modal attention mechanisms require a large amount of multimodal data to train the model so that the model can generalize different sentiment detection tasks and datasets. However, obtaining diverse and representative data can be challenging, especially for domain-specific or sentiment-specific data. 2) Modal imbalance: in cross-modal data, there may be an imbalance between different modalities, i.e., some modalities have less data. This may lead to differences in the performance of the model in processing different modalities, as well as weaker emotion detection for a few modalities. 3) Difficulty in cross-modal alignment: there may be differences and inconsistencies between data in different modalities, e.g., semantic and emotional expressions may be different between text, sound, and video. Therefore, cross-modal attention mechanisms need to address the alignment between modalities to ensure that the model can accurately capture the correlations between different modalities. 4) Model complexity and computational resource requirements: cross-modal attention mechanisms may increase the complexity of the model, including more parameters and computational requirements. This may lead to an increase in time and computational resources for model training and inference, limiting its feasibility and efficiency in real-world applications. 5) Explanation and Interpretability: the complexity of the cross-modal attention mechanism may lead to a decrease in the model's explanation and interpretability. This may make it difficult to understand how the model allocates attention to different modalities and to explain the model's decision-making process and results. 6) Data privacy and security: cross-modal attention mechanisms may need to handle sensitive personal data, such as sound and video data. When using such data for emotion detection, data privacy and security need to be ensured to comply with relevant laws and regulations.

5.3 Future Works

In the task of using cross-modal attention mechanisms to process text, sound, and video data for emotion detection, future directions may include the following: 1) Model design and architecture improvement: further improve the model design and architecture of cross-modal attention mechanisms to enhance the performance and generalization ability of the models. More complex attention mechanisms can be explored, and more modal interactions and alignments can be introduced to better capture the correlations between different modalities. 2) Multimodal datasets and annotation: building larger, diverse, and representative multimodal datasets with accurate sentiment annotation. This will help to improve the generalization ability and performance of the models and advance the research progress in cross-modal sentiment detection. 3) Cross-modal migration learning: using the existing cross-modal data and models, migration learning can be carried out between different tasks and domains. Through transfer learning, existing knowledge and models can be applied to new emotion detection tasks, reducing data requirements and training time. 4) Cross-modal emotion generation: in addition to emotion detection, the task of cross-modal emotion generation can be explored, i.e., generating corresponding emotion expressions based on text, sound and video. This will help to understand and generate multimodal emotional content more comprehensively. 5) Cross-modal emotion understanding applications: apply cross-modal emotion understanding to practical scenarios, such as sentiment analysis, sentiment recommendation, and emotion-driven dialog systems. By combining cross-modal emotion understanding technology with applications in other fields, a richer and smarter human-computer interaction experience can be realized.

6 CONCLUSIONS

This study aims to utilize multimodal data for sentiment analysis by introducing a cross-modal attention mechanism. By constructing text, image and audio feature extractors and using a custom cross-modal attention layer to compute the attention weights between different modalities, we achieve fusion and sentiment classification of multimodal data. Our study achieves good results in emotion detection tasks by introducing a cross-modal attention mechanism. The cross-modal attention mechanism can effectively fuse information from different modalities to improve the performance and accuracy of sentiment analysis.

Our experimental results show that the cross-modal attention mechanism has potential in emotion detection and provides some useful insights for future research. Based on our experimental results, we observed that the cross-modal attention mechanism plays an active role in the emotion detection task. By introducing the attention mechanism, the model is able to better focus on the emotion-related information in different modalities, which improves the accuracy and generalization ability of emotion detection. This study provides some reference value for further development and application in the field of multimodal sentiment analysis.

7 REFERENCES

[1] Bashir, M. F., Javed, A. R., Arshad, M. U., Gadekallu, T. R., Shahzad, W., & Beg, M. O. (2023). Context-aware Emotion Detection from Low-resource Urdu Language Using Deep Neural Network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(5), 1-30. https://doi.org/10.1145/3528576

[2] Baydogan, C. & Alatas, B. (2021). Sentiment Analysis in Social Networks Using Social Spider Optimization Algorithm. *Tehnicki vjesnik-Technical Gazette*, *28*(6), 1943-1951. https://doi.org/10.17559/TV-20200614172445

[3] Kim, E. J. (2022). Applying social computing to analyze the effect of negative emotions on social desirability. *Journal of Logistics, Informatics and Service Science*, *9*(1), 234-257.

[4] Laghari, A. A., He, H., Khan, A., Laghari, R. A., Yin, S., & Wang, J. (2022). Crowdsourcing Platform for QoE Evaluation for Cloud Multimedia Services. *Computer Science and Information Systems*, *19*(3), 1305-1328. https://doi.org/10.2298/CSIS220322038L

[5] Mao, R., Liu, Q., He, K., Li, W., & Cambria, E. (2023). The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, *14*(3), 1743-1753. https://doi.org/10.1109/taffc.2022.3204972

[6] Guo, J. (2022). Deep learning approach to text analysis for

human emotion detection from big data. *Journal of Intelligent Systems*, *31*(1), 113-126.
[7] Wang, X. Y., Yin, Z. X., & Omar, N. A. B. (2023). Evaluation on innovation of emotional marketing in network live broadcasting based on new media environment. *Journal of System and Management Sciences*, *13*(3), 409-420. https://doi.org/10.33168/JSMS.2023.0328
[8] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., & Nanayakkara, V. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*, *81*, 35173-35194. https://doi.org/10.1515/jisys-2022-0001
[9] Lee, S., Lee, S., & Kim, H. (2022). Differential security barriers for virtual emotion detection in maritime transportation stations with cooperative mobile robots and UAVs. *IEEE Transactions on Intelligent Transportation Systems*, *24*(2), 2461-2471. https://doi.org/10.1109/TITS.2022.3172668
[10] Joshi, V. M., Ghongade, R. B., Joshi, A. M., & Kulkarni, R. V. (2022). Deep BiLSTM neural network model for emotion detection using cross-dataset approach. *Biomedical Signal Processing and Control*, *73*, 103407. https://doi.org/10.1016/j.bspc.2021.103407
[11] Veshki, F. G., Ouzir, N., Vorobyov, S. A., & Ollila, E. (2022). Multimodal image fusion via coupled feature learning. *Signal Processing*, *200*, 108637. https://doi.org/10.1016/j.sigpro.2022.108637
[12] Chen, Q., Huang, G., & Wang, Y. (2022). The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, *30*, 2689-2695. https://doi.org/10.1109/TASLP.2022.3192728
[13] Siddiqui, M. F. H., Dhakal, P., Yang, X., & Javaid, A. Y. (2022). A survey on databases for multimodal emotion recognition and an introduction to the VIRI (visible and InfraRed image) database. *Multimodal Technologies and Interaction*, *6*(6), 47. https://doi.org/10.3390/mti6060047
[14] Wang, Y., Xie, Y., Zeng, J., Wang, H., Fan, L., & Song, Y. (2022). Cross-modal fusion for multi-label image classification with attention mechanism. *Computers and Electrical Engineering*, *101*, 108002. https://doi.org/10.1016/j.compeleceng.2022.108002
[15] Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., & Xu, H. (2022). Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, *35*, 26418-26431.
[16] Wen, H., Ding, J., Jin, W., Wang, Y., Xie, Y., & Tang, J. (2022). Graph Neural Networks for multimodal single-cell data integration. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539213
[17] Cheung, T. H. & Lam, K. M. (2022). Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing*, *514*, 1-12. https://doi.org/10.1016/j.neucom.2022.09.140
[18] Guo, Y., Ge, H., & Li, J. (2023). A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism. *Frontiers in Computer Science*, *5*, 1159063. https://doi.org/10.3389/fcomp.2023.1159063
[19] Wu, Y., Guan, X., Zhao, B., Ni, L., & Huang, M. (2023). Vehicle detection based on adaptive multi-modal feature fusion and cross-modal vehicle index using RGB-T images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 8166-8177. https://doi.org/10.1109/JSTARS.2023.3294624
[20] Lee, J.-T., Yun, S., & Jain, M. (2022). Leaky gated cross-attention for weakly supervised multi-modal temporal action localization. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/wacv51458.2022.00089
[21] Ding, C., Sun, S., & Zhao, J. (2023). MST-GAT: A multimodal spatial-temporal graph attention network for time series anomaly detection. *Information Fusion*, *89*, 527-536. https://doi.org/10.1016/j.inffus.2022.08.011
[22] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv 2015. arXiv preprint arXiv:1508.01991.
[23] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, 4489-4497. https://doi.org/10.1109/ICCV.2015.510
[24] Hanan, F. A., Mutalib, S., Yunus, A. M., Rashid, M. F. A., Kamarudin, S. N. K., & Rahman, S. A. (2023). A study on social media responses on road infrastructure using sentiment analysis. *Journal of Logistics, Informatics and Service Science*, *10*(2), 1-14. https://doi.org/10.33168/JLISS.2023.0201
[25] Gavrilă, V., Băjenaru, L., Dobre, C., & Tomescu, M. (2021). Towards the development of a Romanian lexicon for the analysis of emotions in the literary works of canonical authors. *Studies in Informatics and Control*, *30*(2), 111-120. https://doi.org/10.24846/v30i2y202110
[26] Wang, T. (2021). A K-means Group Division and LSTM Based Method for Hotel Demand Forecasting. *Tehnički vjesnik*, *28*(4), 1345-1352. https://doi.org/10.17559/TV-20210507172841
[27] X., C. (2022). Multimedia Teaching System Based on Art Interaction Technology. *Computer Science and Information Systems*, *19*(3), 1517-1532. https://doi.org/10.2298/CSIS220405026C
[28] Jeong, S. & Kim, B. (2021). Network Analysis of Social Awareness of Media Education for Primary School Students Studied through Big Data. *Computer Science and Information Systems*, *18*(2), 575-595. https://doi.org/10.2298/CSIS200316011J

Contact information:

Yuanchen CHAI
(Corresponding author)
Sunshine Ruizhi Securities Consulting (Beijing) Co., LTD 100081
Email: cychhh@126.com