International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n2.10
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Multimodal Emotion Recognition Based on Deep Learning

Jiali Jiang *

Southwest Minzu University, Chengdu, China
* Corresponding Author: Jiali Jiang

ABSTRACT

In recent years, multitask learning-based joint analysis of multiple emotions has emerged as a significant research topic in natural language processing and artificial intelligence. This approach aims to identify multiple emotion categories expressed in discourse by integrating multimodal information and leveraging shared knowledge across related tasks. Sentiment analysis, emotion recognition, and sarcasm detection constitute three closely interconnected tasks in affective computing. This paper focuses on these three tasks - sentiment analysis, emotion recognition, and sarcasm detection - while addressing current challenges in their research. The specific work includes the following three aspects: (1) Due to the limitations of datasets in the development of current Chinese multi-task learning models, this paper establishes a Chinese multi-task multi-modal dialogue emotion corpus to support the development of multi-task multi-modal sentiment analysis. The dataset is annotated with multiple task labels (such as sentiment, emotion, sarcasm, humor, etc.) and, for the first time, manually annotates the correlation between sentiment and emotion, as well as sarcasm and humor. Through scientific evaluation and analysis, it is demonstrated that the dataset possesses high quality and representativeness. (2) Based on the constructed dataset, this paper primarily considers three aspects: context interaction, multi-modal feature fusion, and multi-task learning, and proposes a multi-modal sentiment analysis model based on multi-task learning. Through experimental evaluation, the effectiveness of the model is demonstrated. (3) In response to the model proposed in (2), which fails to consider the interrelationships between tasks, this paper presents a multi-task learning model based on soft parameter sharing to learn the commonalities and differences between different tasks. Experimental results comparing with other advanced baselines demonstrate the innovation and efficiency of the proposed method.

KEYWORDS

Emotional Recognition; Multitasking; Multi-head Attention Mechanism; Neural Networks.

1. INTRODUCTION

Multi-modal sentiment analysis technology strengthens the expression of textual sentiment by establishing the interrelationships between different modalities and utilizing features from non-textual modalities[1]. In recent years, researchers have been dedicated to proposing more precise sentiment analysis models by analyzing the consistency and differences between modalities. For example, Morency et al.[2] pioneered the joint use of visual, audio, and textual features to address the issue of tri-modal sentiment analysis.Pérez-Rosas et al.[3]performed sentiment classification at the discourse level by concatenating features from different modalities and using a support vector machine classifier. With the development of deep neural networks, the introduction of new multi-modal feature fusion techniques has promoted the advancement of multi-modal sentiment analysis. For example, Ghosal et al.[4]used an attention mechanism to learn contextual representations from different modalities and proposed a multi-modal attention framework for discourse-level sentiment

prediction.Kumar et al.[5]performed multi-modal sentiment analysis by modeling the interactions and dependencies between different modalities. Wen et al.[6] designed a cross-modal context gating convolutional network to capture local cross-modal interactions, reduce the impact of irrelevant information, and demonstrated the effectiveness of the proposed method in sentiment analysis tasks through experiments. Inspired by quantum probability, Zhang et al.[7] proposed a density matrix-based framework for fusing text and visual features for dialogue sentiment analysis. Xu et al.[8] built a multi-modal dataset and proposed a multi-interactive memory network (MIMN) based on an attention mechanism to model the relationships between text and images for aspect-level sentiment analysis.

2. RELATED WORK

This chapter mainly introduces the unimodal emotion recognition for speech, text, and vision, including three pre-trained models for each modality and their respective feature extraction algorithms. Additionally, the dataset used in this design is also presented.

2.1. Speech Emotion Recognition

Previous speech emotion recognition primarily used Convolutional Neural Networks (CNN) and LSTM, which overlooked the sequential nature of time. To more effectively capture information in speech signals, this paper proposes a CNN-BiLSTM-based speech emotion recognition model. BiLSTM can better capture the temporal dependencies in speech signals, while CNN is capable of extracting local features from the speech signals. This combination not only improves the accuracy of emotion recognition but also has practical applications in speech emotion analysis.The overall framework of the CNN-BiLSTM model is shown in Figure 1. The model consists of three main parts. First, preprocessing is performed to vectorize the speech signal and extract low-level features from the speech. Next, the second part uses CNN to extract local features from the speech. Then, in the third part, the speech emotion features are input into the BiLSTM layer to capture the temporal dependencies in the speech signal. Finally, the weights are reallocated into the classification layer for the classification output.

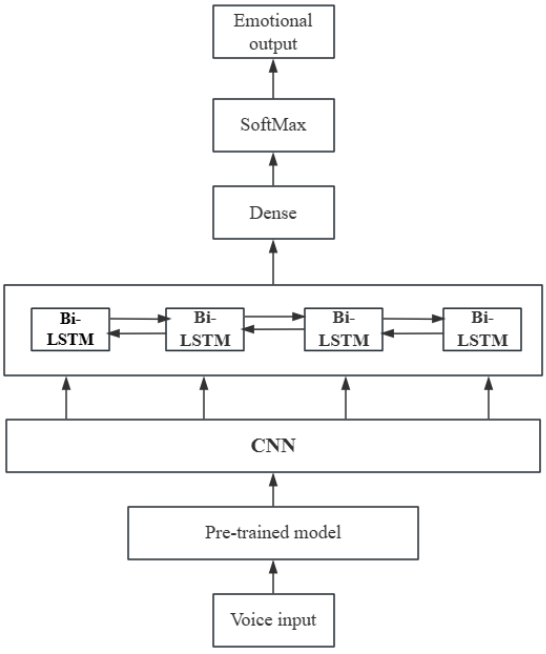

Figure 1. Speech emotion recognition framework

2.2. Text Sentiment Recognition

The LSTM-Attention-based text emotion recognition is a deep learning method commonly used for analyzing and predicting emotions in text (such as positive, negative, or neutral). This method combines LSTM (Long Short-Term Memory networks) and Attention mechanisms, which can effectively capture the long-term dependencies and important information in text for emotion recognition tasks. The process of the entire design is shown in Figure 2.

Figure 2. Text sentiment recognition framework

The text information is represented as vectors through a pre-trained model, and then the attention mechanism dynamically adjusts the weights of different words or features, focusing on the important parts. Next, the LSTM model further extracts feature information, and finally, the emotion is output.

2.3. Visual Emotion Recognition

The CLIP-CNN-LSTM-based image unimodal emotion recognition is a deep learning method that combines CLIP, Convolutional Neural Networks, and Long Short-Term Memory networks. By integrating multimodal feature fusion and temporal modeling, it can more accurately recognize the emotions in images. The entire process is shown in Figure 3.

Figure 3. Video emotion recognition

As shown in Figure 3 above, the video information first passes through the CLIP pre-trained model, which maps the image and text to the same embedding space to extract the semantic features of the image. Then, CNN is used to extract the local features of the image. Next, LSTM is applied to consider the temporal information of the video and capture the temporal dependencies between features. Finally, Softmax is used for emotion output.

For video information, we not only need to consider its image features but also its temporal information. This is because CNN has certain advantages in extracting image features, while LSTM excels in temporal sequence modeling. Therefore, the two models are combined to process video data. The model structure is shown in Figure 4.

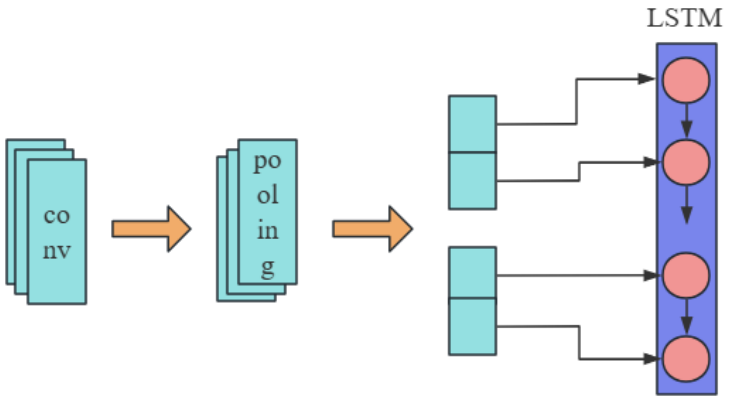

Figure 4. CNN-LSTM flowchart

2.4. Introduction to Datasets

Emotion recognition is a significant challenge in natural language processing and artificial intelligence. As an essential tool for model training and evaluation, datasets corresponding to different scenes and emotional tasks are required by researchers. As shown in Table 1 below, some multimodal emotion recognition datasets are listed.

As people increasingly engage in frequent communication, dialogue-based emotion datasets have been established. IEMOCAP is a dataset recorded by 10 professional actors following a given script and performance, consisting of 10,039 dialogue sentences. These sentences include emotions such as anger, surprise, fear, frustration, happiness, disappointment, excitement, sadness, and neutrality. The dataset mainly includes multimodal data such as dialogue, speech, facial expressions, and gestures. In addition, each dialogue is labeled with emotion categories and emotional intensity.MELDM and MEISD bothEL annotate emotions and sentiment tasks, while CHD-SIMS and ME isIS a ChineseD multim bothodal annotate conversation dataset two tasks established: by emotion a team and sentiment from T.sing CHhua-S University. MIMS isUM aOR Chinese multim isodal a Chinese conversation dataset dataset established containing emotion by and a sentiment team labels from, T but itssinghua drawback University is. that M itUM isOR not is publicly a accessible Chinese. Mem dataset thatotion annot includesates emotion six and different sentiment emotion labels tasks,, but but its its limitation limitation is that is it that it is is not an image publicly accessible-text dataset.. Memotion annotates six different emotion tasks, but its drawback is that it is a text-image dataset.

This paper chooses the CH-SIM Chinese dataset, which, compared to other datasets, focuses on combining text, audio, and visual modalities for emotion analysis. The dataset was released in 2020 by research teams from Tsinghua University and Renmin University of China. Its primary goal is to support multimodal emotion analysis tasks, and it facilitates both single-modal and multimodal emotion analysis research. The dataset contains a total of 2,281 multimodal samples, with data sourced from Chinese TV dramas, variety shows, and interview programs. The emotion categories in the dataset use binary emotion classification (positive/negative) and fine-grained emotion intensity annotations. The emotion intensity for each sample is labeled as a continuous value between 0 and 1, indicating the strength of the emotion. However, this dataset still faces the issue of data imbalance. Therefore, a new dataset has been created by adding a series of video data in video format to address this issue.

3. MULTIMODAL EMOTION RECOGNITION BASED ON MULTI-HEAD ATTENTION MECHANISM

3.1. Monolithic Framework

Multimodal feature fusion refers to combining different modality features to improve the model's performance and robustness. Although feature fusion can better learn the emotional information in the data, it still has the drawback of neglecting the potential interactions and nonlinear relationships between modalities, and capturing interactions between different modalities is challenging. Therefore, this chapter proposes a multimodal fusion based on the attention mechanism, which can dynamically adjust the weights between different features through the complementarity of modalities, capturing the complex relationships between modalities. The overall framework designed is shown in Figure 5 below.

As shown in Figure 5 above, the framework consists of three main parts: feature extraction, multi-head attention, and emotion output. Feature extraction also has three components: speech, text, and visual information, which are extracted using their respective pre-trained models to obtain surface-level features. These features are then further processed by deep learning models to extract deeper features. Next, the extracted features are passed through the multi-head attention mechanism, where

each feature represents Q, K, and V for attention calculation. This adjusts the weights of information from different modalities, highlighting the more important information. Finally, emotion classification is performed to obtain the emotion output.

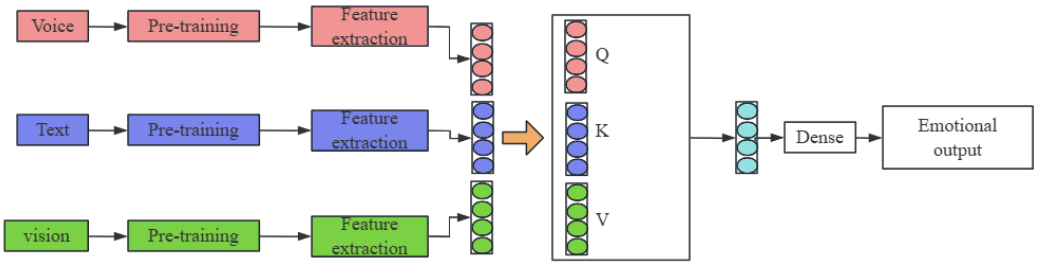

Figure 5. Feature fusion framework

3.2. Single-modal Feature Encoders

Given theGiven provided the image provided, image text,, text and, audio and audio data, we data extract, features we from extract each the modality features using of different each pre modality-trained using models different and pre feature-trained models extraction algorithms and. feature Different extraction feature algorithms encoders. Different are used feature enc tooders extract are the used features to: extract the features.

(1)Speech feature encoder: There are three pre-trained models for speech feature extraction: MFCC, Opensmile, and Wav2vec. After considering various aspects of these models, Wav2vec was ultimately chosen as the pre-trained model for speech in this design. The model's output includes vector representations from multiple hidden layers, with a vector dimension of 128. Feature extraction is done using a CNN-Bi-LSTM model, where a convolutional neural network is first used to extract local features of the speech, and then a bidirectional long short-term memory network is used to preserve the temporal sequence of the speech signal. The final result is the feature vector of the speech.

(2)Text feature encoder: There are two pre-trained models available for text feature extraction: Word2Vec and BERT. Compared to traditional unidirectional models, BERT is a bidirectional model that can consider contextual information, allowing for better understanding of sentence meaning. The model outputs a feature vector with a dimension of 768. Feature extraction uses LSTM and attention mechanisms. The attention mechanism focuses on the important information in the sentence, and then a long short-term memory network is used to capture the dependencies within the sentence, further extracting emotional features. Finally, the feature vector representation of the text is output.

(3)Visual feature encoder: There are two pre-trained models available for visual feature extraction: CLIP and Mediapipe. CLIP can learn textual features along with image data, making it better at extracting visual features. Therefore, the CLIP pre-trained model is chosen. The final output is a vector representation with a dimension of 512. Feature extraction uses a CNN-LSTM combination. Since the data is from video, we need to consider not only the image features but also the temporal information. A convolutional neural network is used to extract image features, while a long short-term memory network is employed to learn temporal features. The final output is the vector representation of the visual information.

3.3. Multi-head Attention Mechanism

The multi-head attention mechanism is built upon the self-attention mechanism by using multiple transformations to generate different Queries, Keys, and Values for calculation. Specifically, it splits the Query, Key, and Value into multiple sub-vectors, computes the attention for each sub-vector, and then concatenates these results, integrating the correlations of the sub-vectors to enhance the effect of the self-attention mechanism. This approach allows the model to focus on different pieces of information simultaneously, improving the model's expressive power. When given the same set of

Queries, Keys, and Values, we hope the model can learn different behavior patterns under a consistent attention mechanism and integrate these behaviors into knowledge, effectively capturing dependencies of various lengths in the sequence, including both short- and long-range dependenciesTherefore, allowing the attention mechanism to combine different subspace representations of Queries, Keys, and Values may have a significant impact on improving the model's performance. We can obtain h sets of different Queries, Keys, and Values (typically h=8) through independent learning, and then feed these h sets of Queries, Keys, and Values in parallel to the attention model. Finally, the outputs of these h attention models are concatenated and transformed through another learnable fully connected layer to produce the final output. This is the basic process of the multi-head attention mechanism.

In multi-head attention, there are h opportunities to learn multiple sets of different Wq, Wk, and Wv, then concatenate the h heads together and perform a final projection. The specific process is shown in Figure 6 below. First, for the same input sequence X, h sets of Wq, Wk, and Wv are defined, and each set computes a different trainable parameter matrix for Query, Key, and Value. These are passed through the self-attention mechanism model to learn, resulting in h different weighted feature matrices Zi (i = 1, 2, ..., h). The Zi matrices are then concatenated into a large feature matrix Z, which is finally transformed through another learnable linear projection model.

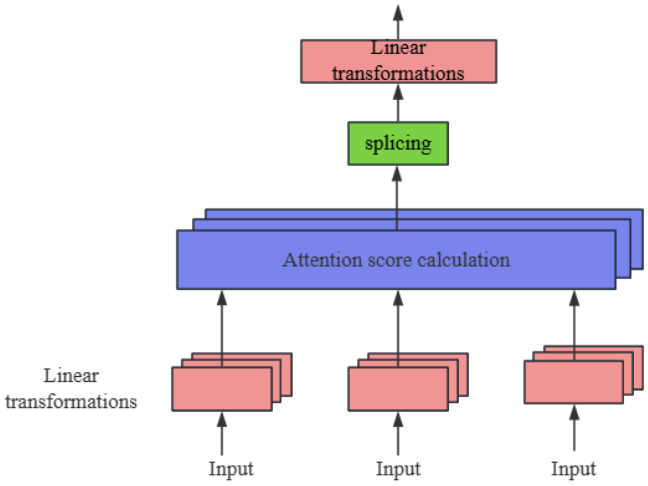

Figure 6. Multi-head attention mechanism

3.4. Cross-modal Feature Fusion

Cross-modal feature fusion is the effective combination of multiple modality information to improve the model's performance on complex tasks. Traditionally, directly concatenating feature vectors from different modalities leads to a dimensional explosion, as the features from different modalities usually have different dimensions. Direct concatenation results in a very large feature vector and may cause overfitting. It can also lead to information redundancy, increasing computational complexity and reducing model efficiency. Additionally, it may overlook the potential interdependencies between modalities, leading to missing information. Therefore, this section proposes a cross-modal fusion based on the multi-head attention mechanism, which adjusts the information from each modality according to attention scores to highlight important information. It can capture fine-grained interactions, enhancing interpretability. The features of speech, text, and vision modalities are extracted to obtain feature vectors for the three modalities h_a, h_t, h_v .The specific splicing method is as follows:

First, feature projection is performed for linear mapping according to the formula above. Then, multi-head splitting is applied, where Q, K, and V are split into h heads, each with a dimension of d/h. After that, cross-modal attention computation is performed based on the following formula.

$$head_i = Soft\max(\frac{QK^T}{\sqrt{d/h}})V$$

Attention between pairs can be obtained, for example, the attention between image and text is computed by first mapping them to Q, K, and V through a linear layer, and then calculating using the formula above. Finally, the three heads are mapped through a fully connected layer, and the fused feature vector is obtained according to the formula above.

4. MULTIMODAL EMOTION RECOGNITION FOR MULTI-TASK INTERACTIVE LEARNING

Currently, multi-task learning-based joint analysis of multiple sentiment tasks is a widely used technique in the field of natural language processing. The main idea is to combine multiple related tasks such as sentiment analysis, emotion recognition, and sarcasm detection, to improve the performance of different tasks by sharing underlying parameters. In recent years, multi-task learning models primarily adopt a hard parameter sharing structure. However, the models proposed under this paradigm only focus on modeling the correlations between tasks, while ignoring the heterogeneity between them.

4.1. Multitasking Model Framework

The model framework presented in this chapter consists of three main modules：

(1)Multimodal feature extraction: The text, visual, and audio content of the target utterance are separately processed through three pre-trained models, BERT, ResNet, and VGGish, to obtain the initial text, visual, and acoustic features, which serve as the input for the multimodal fusion stage.

(2)Multimodal feature fusion: A text-centered attention mechanism fusion strategy is proposed, which can automatically learn the alignment between text and image, as well as text and speech. In this approach, the text is treated as the query vector (Query), while the image and speech are considered the key vector (Key) and value vector (Value), respectively. This setup forces the model to learn the appropriate attention weights for the different modal representations in the related image and speech, in order to obtain the text-visual and text-acoustic fused representations. Then, these two bimodal fused representations are concatenated and passed through a set of self-attention mechanism layers to obtain the final multimodal fused representation.

(3)Multitask interaction: The multimodal fused representation is input into a multitask learning network, which consists of a set of gating networks, three fully connected layers (i.e., FC1, FC2, FC3), and three GRU classification branches. The fully connected layers (FCs) generate different feature representations of the same dimension, learning the representations of specific regions for different target tasks. A gating mixture network is introduced to assign values to the output based on the tasks. By combining the weighted outputs, automatic parameter allocation can be achieved, aiming to capture the shared information across multiple tasks without the need to add many new parameters for each individual task.

4.2. Multimodal Feature Fusion

In the feature fusion stage, a text-centered bimodal fusion method is first set up. Two cross-modal attention mechanisms are used to fuse text with visual and text with audio features, obtaining the text-visual fused features and text-acoustic fused features. Finally, the bimodal fused features are concatenated and passed through a self-attention layer to obtain the multimodal fused features, as shown in Figure 7.

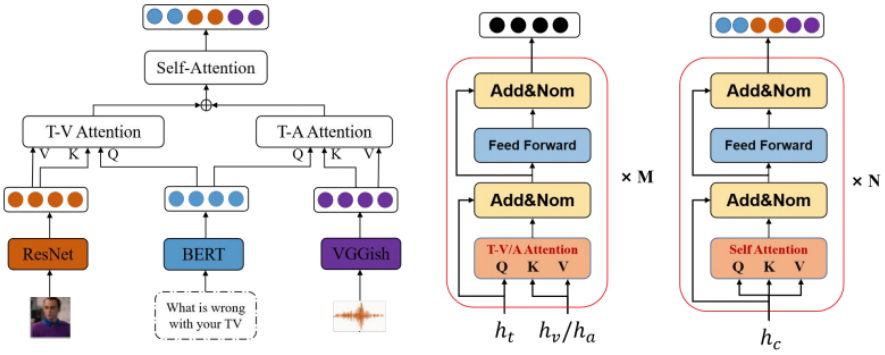

Figure 7. Feature fusion method

4.3. Multi-task Interactive Learning

The essence of multitask learning is shared representations. Inspired by MMOE, this chapter introduces a multitask learning paradigm based on soft parameter sharing for sentiment analysis, emotion recognition, and sarcasm detection, and develops an interactive learning network consisting of gating units and three fully connected layers (FCs), as shown in Figure 8. Specifically, each fully connected layer (FC) is used to generate different feature representations of the same dimension, learning specific region representations for different tasks. The model sets a gating network (Gating network) for each target task, and all target tasks share the outputs from multiple fully connected layers. Each fully connected layer has its own specialized learning direction, with each gating network learning to select the signal weight for the output of each fully connected layer. Finally, the outputs of each fully connected layer are weighted and combined based on the gating network's output weights, and applied to the sentiment, emotion, and sarcasm tasks respectively. The method is as follows:

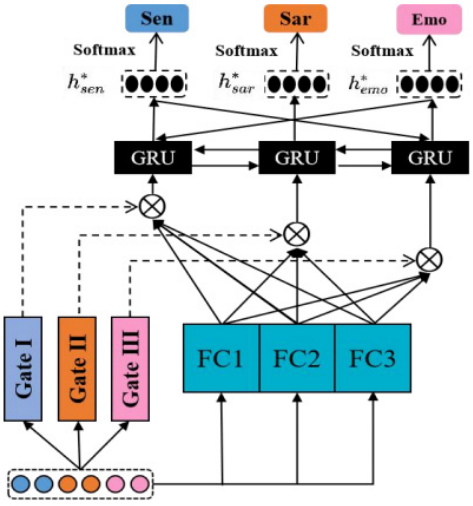

Figure 8. Multi-task interactive learning model

5. EXPERIMENTS

5.1. Experimental Parameter Settings

The experiment uses the Python programming language, with pretraining utilizing the PyTorch deep learning framework, and feature extraction using the TensorFlow deep learning framework. The GPU used is the NVIDIA GeForce RTX 4060. The extracted audio and text feature sizes are 128-dimensional and 768-dimensional, respectively. The optimizer used is Adam, the activation function

is ReLU, the loss function is SparseCategoricalCrossentropy, the learning rate is 0.001, and the sample training batch size is 16. The detailed parameters are shown in the table below.

Table 1. Basic parameters of the experiment

parameter	value
Optimizer	Adam
Learning rate	0.001
Batch size	16
Activate the function	ReLU
Loss function	SparseCategoricalCrossentropy

5.2. Experimental Results

First, a comparison was made between unimodal and multimodal emotion recognition. The table below shows the average accuracy results for emotion recognition.

Table 2. Comparison of single-modal and multimodal results

modality	model	Identify the results
Unimodal voice	Bi-LSTM	51.20
	CNN-Bi-LSTM	56.02
	CNN-Bi-LSTM-Attention	54.27
Unimodal text	LSTM	60.83
	Attention-LSTM	61.27
Unimodal vision	CNN	61.27
	CNN-LSTM	64.77
Three-modal	Long attention	68.23
	Multitasking federation	75.11

From the experimental results, it can be seen that the multimodal emotion recognition results are significantly higher than the unimodal results. After introducing information from other modalities, the experimental setup for multimodal recognition shows that integrating additional modal information helps improve the accuracy of various emotion tasks. It is also evident from the experimental results that the proposed model structure greatly enhances the results of emotion recognition.

6. CONCLUSION

Multitask learning-based joint analysis of sentiment, emotion, and sarcasm is a complex task that involves natural language processing and deep learning. Multimodal joint analysis of sentiment, emotion, and sarcasm involves various types of data, such as text, images, audio, etc. Effectively combining these different types of data to improve the model's performance is a significant challenge. Additionally, in sentiment, emotion, and sarcasm joint analysis, there may be interdependencies and interactions between these tasks, and handling the relationships between different tasks is a tricky problem faced when building the model. To address the above challenges, the innovative work of this paper is summarized as follows:

(1) Due to the lack of a Chinese multimodal dialogue dataset with multiple emotion annotations, this paper proposes a Chinese multimodal multi-emotion dialogue dataset to fill this gap. Compared to previous datasets, this dataset manually annotates the correlations between different emotions for the

first time, facilitating the subsequent development of multitask models. Additionally, the dataset is collected from multi-party dialogue video clips, and it also includes annotations of the speakers' personal information and the topics of the dialogues. The dataset takes into account various factors that influence the emotional trends in human communication.

(2) This paper, with the background of multimodal dialogue sentiment analysis, comprehensively considers context information, modal feature information, and related task information. It proposes a multitask learning model to verify the quality of the constructed dataset. Through comparative experiments, the practicality and effectiveness of the proposed method are demonstrated.

(3) This paper argues that sentiment, sarcasm, and emotion are closely related and proposes a multitask interactive learning framework based on soft parameter sharing for joint detection of sentiment, emotion, and sarcasm. The main idea is to automatically learn the alignment between text and images, as well as between text and audio, and to learn both common and task-specific knowledge from related tasks. Finally, extensive experiments on two benchmark datasets demonstrate the effectiveness of the proposed model, showing significant improvements over the state-of-the-art baselines.

REFERENCES

[1] Middya A I, Nag B, Roy S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities[J]. Knowledge-Based Systems, 2022, 244: 108580.

[2] Morency L-P, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web[C]// Proceedings of the 13th international conference on multimodal interfaces, 2011: 169-176.

[3] Pérez-Rosas V, Mihalcea R, Morency L P. Utterance-Level Multimodal Sentiment Analysis[C]// Association for Computational Linguistics. ACL, 2013.

[4] Ghosal D, Akhtar M S, Chauhan D, et al. Contextual inter-modal attention for multi- modal sentiment analysis[C]// Proceedings of the 2018 conference on empirical methods in natural language processing, 2018: 3454-3466.

[5] Kumar A, Vepa J. Gated mechanism for attention based multi modal sentiment analysis[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 4477-4481.

[6] Wen H, You S, Fu Y. Cross-modal context-gated convolution for multi-modal sentiment analysis[J]. Pattern Recognition Letters, 2021, 146: 252-259.

[7] Zhang Y, Song D, Zhang P, et al. A Quantum-Inspired Multimodal Sentiment Analysis Framework[J]. Theoretical Computer ence, 2018: 21-40.

[8] Xu N, Mao W, Chen G. Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis[C]// The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 2019.

[9] Chuang Z-J, Wu C-H. Multi-modal emotion recognition from speech and text[C]// International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing, 2004: 45-62.

[10] Datcu D, Rothkrantz L J. Semantic audiovisual data fusion for automatic emotion recognition[J]. Emotion recognition: a pattern analysis approach, 2015: 411-435.

[11] Poria S, Hazarika D, Majumder N, et al. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations[J], 2018.

[12] Majumder N, Poria S, Hazarika D, et al. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations [C] // 2019: 6818-6825.

[13] Zhang Y, Li Q, Song D, et al. Quantum-Inspired Interactive Networks for Conversational Sentiment Analysis[C]// Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}, 2019.