The Performance of a Temporal Multi-Modal Sentiment Analysis Model Based on Multitask Learning in Social Networks

Lin He, Haili Lu*
Faculty of Education, Shaanxi Normal University, Xi'an, 710062, China

***Abstract*—When conducting sentiment analysis on social networks, facing the challenge of temporal and multi-modal data, it is necessary to enable the model to deeply mine and combine information from various modalities. Therefore, this study constructs an emotion analysis model based on multitask learning. This model utilizes a comprehensive framework of convolutional networks, bidirectional gated recurrent units, and multi head self-attention mechanisms to represent single modal temporal features in an innovative way, and adopts a cross modal feature fusion strategy. The experiment showed that the model accomplished 0.83 average precision and a 0.83 F1-value, respectively. In contrast with multi-scale attention (0.69, 0.70), aspect-based sentiment analysis (0.78, 0.74), and long short-term memory network (0.71, 0.78) models, this model demonstrated higher robustness and classification accuracy. Especially in terms of parallel computing efficiency, the acceleration ratio of the model reached 1.61, which is the highest among all compared models, highlighting the potential for time savings in large data volumes. This study has shown good performance in sentiment analysis in social networks, providing a novel perspective for solving complex sentiment classification problems.**

***Keywords*—Multi task learning; multi-modal; emotional analysis; attention mechanism; feature fusion**

I. Introduction

Due to the rapid growth of social networks and the increasing number of users expressing emotions on platforms, social media has become an indispensable part of people's daily lives [1]. Users can not only communicate conveniently on these platforms, but also share emotions and viewpoints, which has become one of the main channels for social emotional and opinion dissemination. Users express their emotions and opinions through publishing text, images, videos, and other forms of content, thus forming a vast and rich information network. Therefore, analyzing and understanding emotions and perspectives on social media is of great significance for grasping social emotional dynamics and gaining a deeper understanding of user needs [2]. However, people's emotional expression on platforms shows a trend of diversification and complexity, which puts higher demands on sentiment analysis technology.

Emotional analysis requires a deep understanding and modeling of user emotional states, in order to extract relevant emotional features from complex data [3]. Existing sentiment analysis tools mostly focus on text content and rarely involve emotion recognition in images or videos, especially lacking sentiment flow analysis in time series [4]. In addition, existing sentiment analysis techniques face the problem of difficulty in accurately capturing and analyzing emotional information in multi-modal data containing temporal information. Therefore, new sentiment analysis techniques urgently need to be proposed to address the aforementioned challenges.

Multimodal data analysis includes various forms of data such as text, images, videos, etc. It is suitable for analyzing complex content in social media and can provide a more comprehensive understanding of user emotional states and dynamic changes [5]. However, there are still some challenges in this analysis method, such as how to effectively integrate data from different modalities, and how to accurately capture the impact of temporal information on emotions. In addition, due to the diversity and complexity of data, sentiment analysis models often find it difficult to fully consider all possible scenarios and contexts, which affects the accuracy and stability of analysis results. Therefore, further research and improvement of algorithms are needed to enhance the accuracy of sentiment analysis in multimodal temporal data.

In view of this, in order to deeply explore and accurately analyze emotions in social media, this paper proposes a multi-task learning emotion analysis model combining multi-modal data and temporal characteristics. The innovation of the research lies in applying multimodal fusion and time series analysis to emotion recognition tasks to improve the comprehensiveness and effectiveness of the model, aiming to achieve public opinion monitoring and emotional recommendation.

The contribution of the research is to fill the gap in multimodal temporal data processing with existing sentiment analysis techniques, providing new ideas and methods for the development of social media sentiment analysis. By delving into emotional information in multimodal data and combining it with time series analysis, research is expected to provide more accurate and comprehensive emotional analysis services for social media platforms, thereby promoting the intelligent development of social media.

The research content contains six sections. Section II is an overview of the current research status of Temporal Multi-modal Sentiment Analysis Models (TMSAM) for multitasking learning. Section III introduces a single modal temporal feature representation method. It fully explores the intrinsic temporal information in sequence data through convolutional networks (CNN), bidirectional Gated Recycle Unit (BiGRU), and multi-head self attention mechanisms (MH-SAM), and proposes a cross modal feature fusion method. Section IV gives details about the application of sentiment analysis. Result and discussion is given in Section V. Finally, Section VI concludes

the paper.

II. Related Works

Emotional analysis has always been a highly focused research field in social networks. The focus of research on sentiment analysis models for multi-modal data and temporal features mainly includes the fusion of sentiment features, modeling of temporal information, and the application of multitasking learning. Rahmani et al. designed a multi-modal emotion prediction model based on a cognitive perception framework. This model constructed an adaptive tree by hierarchical partitioning of users, and then trained sub models of Long Short Term Memory (LSTM), utilizing attention-based fusion to transfer cognitive oriented-knowledge within the tree. This algorithm could better use potential clues and promote prediction results compared to other ensemble methods [6]. Middya et al. explored various fusion strategies, including early fusion, late fusion, and attention mechanisms, to effectively combine and utilize complementary data from diverse modalities [7]. Zhou et al. established a new multi-modal model for audio-visual emotion recognition built on adaptive multi-level factor decomposition bilinear pooling. This model utilized FCN networks to recognize speech emotions and adopted adaptive strategies to calculate fusion weights. Compared to other methods, this method outperformed current advanced data with 71.40% accuracy [8]. Zhang et al. proposed a cross modal semantic content association method. It took pre trained CNN to encode the content of visual sub regions, then associated them with images, and used CASR networks to process class aware statements, finally feeding them back to within class dependency LSTM. The proposed correlation method has been proven to be effective [9].

In addition, the exploration of MTL methods has become a prominent research focus to optimize the sentiment analysis models in social networks. Kumari R et al. jointly learned freshness and emotion error detection from target text and proposed a MTL based emotion recognition and error detection model. The proposed model has improved accuracy compared to other models on four different datasets [10]. Akhtar et al. utilized the correlation between participating tasks in a multitasking framework and set three different settings. Each setting includes two tasks: emotion classification and emotion intensity. The evaluation showed that this framework produced better performance compared to single task learning frameworks [11]. Plaza Del Arco et al. utilized shared emotional knowledge and Transformer models to detect various hate speech in social media networks. By jointly learning multiple related tasks, such as sentiment polarity classification, sentiment recognition, and subjective detection, MTL utilized shared representations and promoted the extraction of task specific features, thereby improving the model's generalization ability and adaptability. The combination helped to more accurately detect hate speech across datasets when multitasking [12].

In summary, at present, emotional analysis in social networks urgently requires in-depth research on multi-modal data and time series. Current research mostly addresses the complexity of sentiment analysis by building MTL frameworks, while integrating different modalities of data processing cannot fully understand user emotions. On this basis, this study proposes a TMSAM based on MTL to address emotional complexity to conduct targeted analysis of emotions in social networks.

III. Multi-modal Sentiment Analysis Model and Experimental Design

This study constructs an emotion analysis model based on multitask learning. It utilizes a comprehensive framework of CNN, BiGRU, and MH-SAM to represent single modal temporal features in an innovative way, and adopts a cross modal feature fusion strategy.

A. Design of TMSAM Model Based on MTL

MTL is inspired by human inductive learning, which improves the generalization performance of models by simultaneously learning information from multiple related tasks and achieving information sharing. Multi-modal sentiment analysis based on multitask learning faces three major challenges: intra modal, out modal, and inter modal interactions [13]. Inter modal interaction involves the fusion of multiple modal features. Intra modal interaction involves contextual interaction of the target discourse. Out modal involves the correlation and influence between different emotional tasks. This study proposes a sentiment analysis model based on deep multitasking learning from these three aspects. This model combines sentiment and emotion analysis, utilizes BiGRU to capture contextual information of conversations, and achieves inter modal interaction through attention mechanisms, while predicting emotions and emotions. The model structure is Fig. 1.

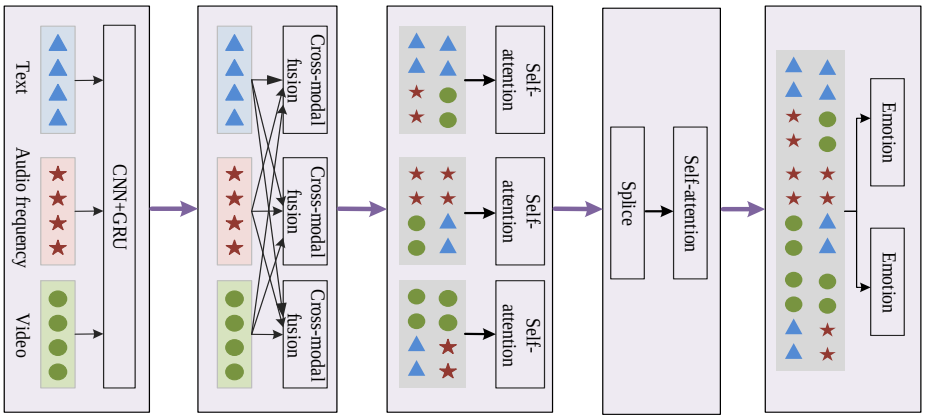


Fig. 1. Multi task sentiment analysis model.

In Fig. 1, the first step is intra modal feature extraction, where each conversation segment records the language text, facial expressions, and audio information of different speakers in chronological order. Due to the changing emotions during the dialogue process, discourse emotion detection is dependent on its context [14]. Therefore, this study uses pre trained models to extract and contextual features of the target discourse based on the three modalities of text, video, and audio, and concatenates them to represent the final features of each modality. The subsequent stage involves multi-modal feature fusion and MTL, using two different fully connected sub-networks to classify emotions and emotions in the obtained feature matrix [15].

In the step of intra modal feature extraction, this study conducted single modal feature extraction. Assuming each sample $X=(x_1,x_2,\cdots,x_n)$ in the dataset be a time series of length L. This time series consists of n segments of dialogue, text, video, and audio. Feature extraction utilizes CNN and BiGRU to obtain global contextual feature information to extract internal features of a single modality [16]. This study used a set of CNNs with the same width, height, and sequence dimensions to extract local information, as shown in Fig. 2.

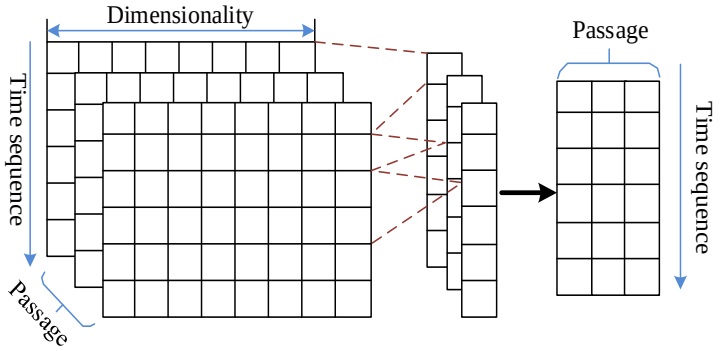


Fig. 2. Process of CNN extracting local temporal information.

In Fig. 2, this study uses a set of convolution kernels with the same height, width, and sequence dimension to extract local information. The data processed by CNN maintains a time series structure, but the dimension is unified as the number of convolution kernels $d=(k\in\{T,A,V\})$. When the stride of the convolutional kernel is 1 and no padding is used, the original time series length is shortened, which helps to accelerate subsequent recurrent neural network training and reduce the shape of the attention matrix [17]. Text features are extracted through GloVe. Audio features are extracted through CoVarRep. Video features are extracted through Facet. The obtained features are averaged based on word dimensions to obtain sentence level feature representations, as shown in Eq. (1).

$$X^m=\left[X_1^m,X_1^m,\cdots,X_n^m\right]\in R^{T\times d_m} \tag{1}$$

In Eq. (1), $m\in\{T,A,V\}$, T, V, and A are text, video, and audio formats. The sampling rate of different modal features is different, and the dimension feature $d\in\{T,A,V\}$ and sequence length $L\in\{T,A,V\}$ are different. Using CNN as a sequence alignment tool in multi-modal sentiment analysis, similar to a fully connected layer (FCL), passes the input sequence to a 1D convolutional layer. The expression is Eq. (2).

$$\hat{X}_{|T,A,V|}=Conv1D\left(X_{|T,A,V|},X_{|T,A,V|}\right)\in R^{T\times d_{|T,A,V|}} \tag{2}$$

In Eq. (2), L represents a time series of length L. $k_{|T,A,V|}$ represents modality's convolution kernel size. BiGRU consists of the update/reset gate, with a simple structure that can alleviate the issues of gradient dispersion and explosion [18-19], as shown in Eq. (3).

$$\begin{cases} r_j=\delta\left(U_r x_{ij}+W_r h_{j-i}+b_r\right) \\ z_j=\delta\left(U_z x_{ij}+W_z h_{j-i}+b_z\right) \end{cases} \tag{3}$$

In Eq. (3), x_{ij} represents the input characteristic value of the j-th element in sample i. U,W is weight. b means the bias coefficient. The hidden layer state of the modal sequence is Eq. (4).

$$\begin{cases} \hat{h}_t=\tanh\left(U_h x_{ij}+W_h h_{j-i}+b_h\right) \\ h_t=z_t+\hat{h}_t+(1-z_t)h_{t-1} \end{cases} \tag{4}$$

In Eq. (4), h_t means the hidden layer state of the modal sequence at time t. Continuing to input the data processed by CNN into BiGRU, continuously update the hidden state, and extract the bidirectional hidden state corresponding to the time series as high-order time features. Its expression is Eq. (5).

$$Z_{\{T,A,V\}}=BiGRU\left(X_{\{T,A,V\}}\right)\in R^{L\times d} \tag{5}$$

After obtaining text, visual, and audio features, multi-modal feature fusion is then performed. This process integrates the feature information from different modes or sensors to lift the robustness of the model. The core idea is to combine information from different modalities to obtain more comprehensive and accurate information. Common multi-modal feature fusion ways include early/late/hybrid fusion strategies [20].

B. Establishment of Feature Fusion and Performance Evaluation Methods in Multi-Modal Sentiment Analysis

In multi-modal emotion classification, the importance of each modality varies for different tasks, sometimes through

facial expressions, and sometimes through language expression. Therefore, the contribution of each modality is crucial to the final classification result. Cross modal attention can capture the connection between modalities and achieve dynamic interaction. MH-SAM can reduce dependence on external data and is beneficial for capturing internal connections of data or features [21]. The model obtains the dependency relationships between words by analyzing the dependency tree of sentences. Fig. 3 shows the dependency relationships of sentences.

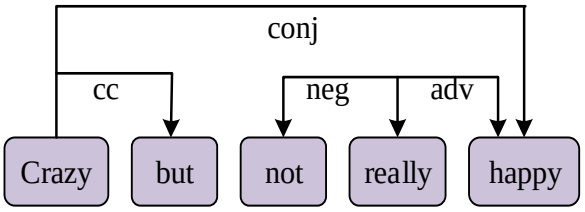


Fig. 3. Dependency relationships between sentences.

After obtaining the dependency relationships between words through the semantic dependency tree of sentences, the model uses bidirectional LSTM to extract sentence representations from text data, and then uses CNN combined with dependency relationships to encode the sentence representations to obtain node representations. Then, using attention mechanism, the node representation is reassigned to the emotional weights of the sentence representation and inputted into the FCL. Finally, the sentiment orientation of the sentence is determined through a discriminator. The multi head attention mechanism is Fig. 4.

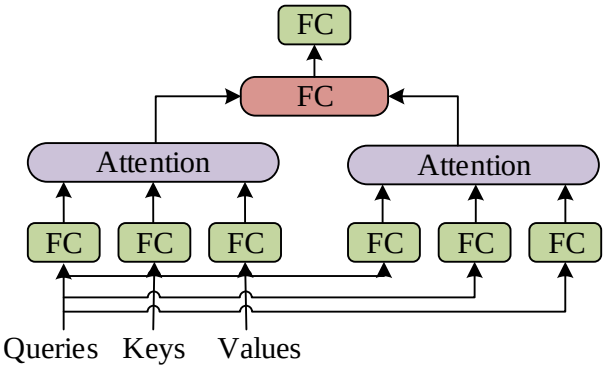


Fig. 4. Schematic diagram of multi head attention mechanism.

In Fig. 4, this study combines the advantages of cross modal attention and multi head self attention (MHSA) and proposes a multi-level cross modal feature fusion method. This method allows the attention mechanism to learn different behaviors based on the same set of queries, keys, and values, and allows the attention mechanism to combine different subspace representations to transform values, keys, and queries, which may be beneficial. Multiple attention pooled outputs are connected together and transformed through a learned linear projection to generate the final output. This is called multi-head attention. The calculation formula for MHSA from modality to modality is Eq. (6).

$$\begin{cases} Y_\lambda = CM_{\eta\to\lambda}\left(X_\lambda, X_\eta\right) = soft\max\left(\dfrac{Q_\lambda K_\eta^L}{\sqrt{d_k}}\right)V_\eta \\ Y_\lambda = soft\max\left(\dfrac{X_\lambda W_{Q_\lambda} W_{k\eta}^L X_\eta^L}{\sqrt{d_k}}\right)X_\eta W_\eta \end{cases} \tag{6}$$

In Eq. (6), $\sqrt{d_k}$ represents the scaling factor. Q_λ, K_η and V_η are the query, key and value vectors. Operation QK^L can obtain the attention weight matrix, and the specific calculation formula for the three vectors is Eq. (7).

$$\begin{cases} Q_\lambda = X_\lambda W_{Q_\lambda}, W_{Q_\lambda} \in R^{d_\lambda \times d_K} \\ K_\eta = X_\eta W_{K_\eta}, W_{K_\eta} \in R^{d_\eta \times d_k} \\ V_\eta = X_\eta W_{V_\eta}, W_{V_\eta} \in R^{d_\eta \times d_k} \end{cases} \tag{7}$$

In Eq. (7), W_{Q_λ}, W_{K_η}, and W_{V_η} are the mapping matrices of the query/key/value vector. This study utilizes a cross modal attention mechanism to fuse features pairs by pairs between different modalities, capturing the correlation between modalities. This stage is called the cross modal feature fusion layer [22]. Then, the obtained pairwise fused feature matrix is concatenated and the internal correlation of modal features is captured through self attention mechanism. Furthermore, these modal feature matrices are concatenated twice and fused again through self attention to capture the correlation between different modal characteristics and identify the modal information that contributes the most to the recognition task. The data is mapped to a low dimensional space, and the outputs of all attention heads are gathered to obtain the complete output result, as shown in Eq. (8).

$$\begin{cases} Z_Z = \left[\hat{Z}_T \oplus \hat{Z}_A \oplus \hat{Z}_V\right] \\ Z_I = attention\left(Z_Z\right) \end{cases} \tag{8}$$

In Eq. (8), $\hat{Z}_T$, $\hat{Z}_A$, and $\hat{Z}_V$ represent the first fusion feature of text, audio, and video. $\oplus$ represents splicing operation. Z_Z represents the secondary fusion feature of the sample. Z_I represents the final fusion feature of the modality. After obtaining multi-level modal temporal feature fusion, to perform MTL and predict the probability of different label categories for each emotion task. The prediction process is Fig. 5.

In Fig. 5, this study uses a feature matrix generated by cross modal feature fusion, which is processed through three FCLs for sentiment classification tasks. At the same time, two different fully connected sub networks are used to classify the sentiment of the feature matrix. MTL is the process of improving the generalization ability of multiple tasks by sharing underlying representations. This study uses multi-modal fusion feature Z_I as a hard parameter for sharing, and uses a FCL to obtain the predicted probabilities of different label categories for each emotion task.

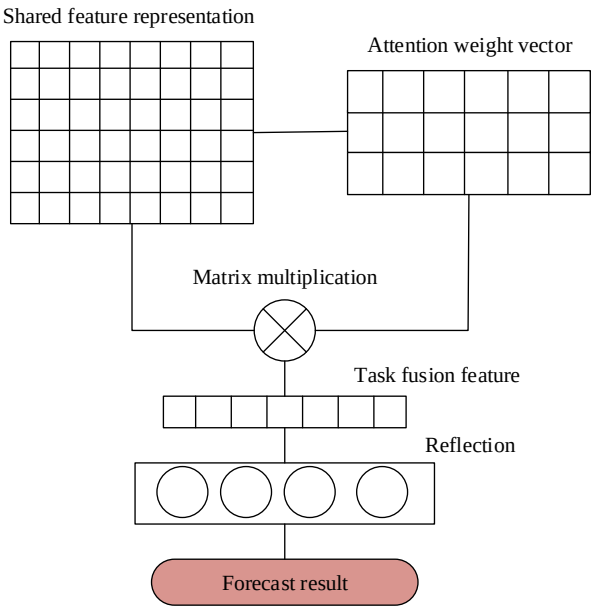


Fig. 5. Emotional task prediction flowchart.

The predicted probability is Eq. (9).

$$Y_k = soft\max\left(W_k Z_I + b_k\right) \tag{9}$$

In Eq. (9), k represents different tasks. W_k represents the weight parameter. b_k represents the bias term. The network is trained taking a cross entropy loss function, as shown in Eq. (10).

$$Loss_k = -\sum_{i=1}^{D}\sum_{j=1}^{C_k} \hat{y}_i^k \log\left(y_i^k\right) + \alpha\theta^2 \tag{10}$$

In Eq. (10), D is the quantity of training samples. C_k represents the amount of different task categories. $\hat{y}_i^k$ and y_i^k represent the true predicted categories and predicted categories for different tasks. $\alpha\left\|\theta^2\right\|$ represents the regularization function. To better evaluate the performance of TMSAM in social networks, this study selected macro F1 score, precision, recall, and acceleration ratio as the evaluation indicators for the experiment. The calculation formula for accuracy is Eq. (11).

$$precision = \frac{TP}{TP + FP} \tag{11}$$

In Eq. (11), TP and FP represent the number of correctly and incorrectly predicted positive emotion words. Precision tests the sample numbers predicted by the model as positive examples are true examples. The recall rate measures the proportion of real cases, and the calculation is Eq. (12).

$$recall = \frac{TP}{TP + FN} \tag{12}$$

In Eq. (12), FN means the incorrectly predicted negative emotion words. The F1 score combines recall and precision to evaluate the classification models, the formula is Eq. (13).

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{13}$$

In Eq. (13), the higher the F1 score, the better the model performance. In terms of high-performance computing, acceleration ratio refers to the ratio of serial to parallel program execution time. It is used to measure the degree of performance improvement of parallel computing compared to serial computing. The calculation of acceleration ratio is Eq. (14).

$$S = \frac{T_{cpu}}{T_{gpu}} \tag{14}$$

In Eq. (14), T_{cpu} and T_{gpu} represent the time it takes for the model to run an epoch on both CPU and GPU. If the acceleration ratio is greater than 1, it indicates that parallel computing is more efficient than serial computing.

IV. The Application of Sentiment Analysis Models in Social Network Datasets

This paper studies the performance of the TMSAM model built on MTL in social networks, with a particular focus on the dataset and parameter settings applicable to social network data. To assess the sentiment analysis models, this study selected six real-world social network datasets containing a large amount of social media text and visual data. These sentiment analysis datasets are Twitter, Facebook, Reddit, Weibo, Instagram, and YouTube. Six datasets cover various emotional categories and complex social interaction information, providing a challenging testing platform. Table I provides specific information for each dataset.

The dataset constructed in Table I is segmented into three categories: training, testing, and validation sets. Table II shows the partitioned dataset information.

TABLE I. Multi-modal Sentiment Analysis Dataset

Dataset	Type	Size	Modality
Twitter	Text and Dialogue	10145 sentences	T/I/A
Facebook	Dialogue and comments	14879 sentences	T/I/A
Reddit	Dialogue and comments	21532 sentences	T/I/A
weibo	Text and video	12367 sentences	T/I/A
Instagram	Images and Text	9856 sentences	T/I/A
YouTube	Videos and comments	14623 sentences	T/I/A

Note: T/I/A represents the text, image, audio.

After determining the number of three sets in Table II, experimental parameters need to be set. This experiment was written in Python 3.7, using a deep learning framework of Python 1.2.0 and a graphics card of TelsaK80. This study used the Python framework in deep learning for encoding. To prevent over-fitting, stop training when the model's performance on the validation set begins to decline. Table III shows the parameters for model training.

To ensure optimal performance of the model in social network environments, careful parameter tuning was carried out. To better demonstrate the impact of attention mechanism on GCN, text embeddings on the dataset were visualized, and the specific results are exhibited in Fig. 6.

TABLE II. PARTITIONED DATASET

Task	Classification	Training set	Verification set	Test set
Emotion	negative	2616	805	793
	neutral	9247	2703	2258
	positive	2177	662	934
Mood	happy	1982	425	193
	detest	832	129	28
	sad	631	179	115
	frightened	488	103	64
	surprise	545	523	386
	angry	1608	384	371
	neutral	7654	2261	2754

According to Fig. 6, the colors of the dots represent different emotional labels. After introducing attention mechanism, the model learned the features of text embedding better, resulting in more separability of samples in the reduced subspace. To better evaluate the research model performance in social networks, this study selected F1 score and precision as the evaluation indicators for the experiment. This study selected the ABSA model from study [5], the LSTM model from reference [14], and the MSA model from study [25] for comparative analysis with the research model TMSAM. The accuracy experimental results of the four models are displayed in Fig. 7.

In Fig. 7, the performance of the four models on six social network datasets has their own advantages and disadvantages. The research model performs the best, with the highest mean accuracy (MA) in each dataset, at 0.83. The ABSA model and LSTM model performs moderately, with an MA of 0.78 and 0.71 in six datasets, respectively. The MSA model performs the worst, with an MA of only 0.69. The results of comparing the F1 values of the four models using the same method are shown in Fig. 8.

TABLE III. SPECIFIC PARAMETERS

Parameter	Value	Parameter	Value
GRU hidden layer dimension	200	optimizer	Adam
batch size	32	optimization function	Binary Cross Entropy
learning rate	0.001	activation function	ReLU
dropout	0.3	L2 regularization parameters	Le-5
First level attention dimension/number of heads	400/6	The number of layers in GCN	2
Second level attention dimension/number of heads	1200/8	Epochs	20
Cross modal attention dimension/head count	200/4	Word vector dimension d	300

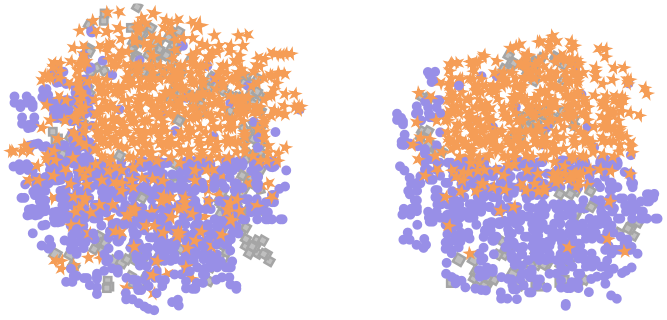
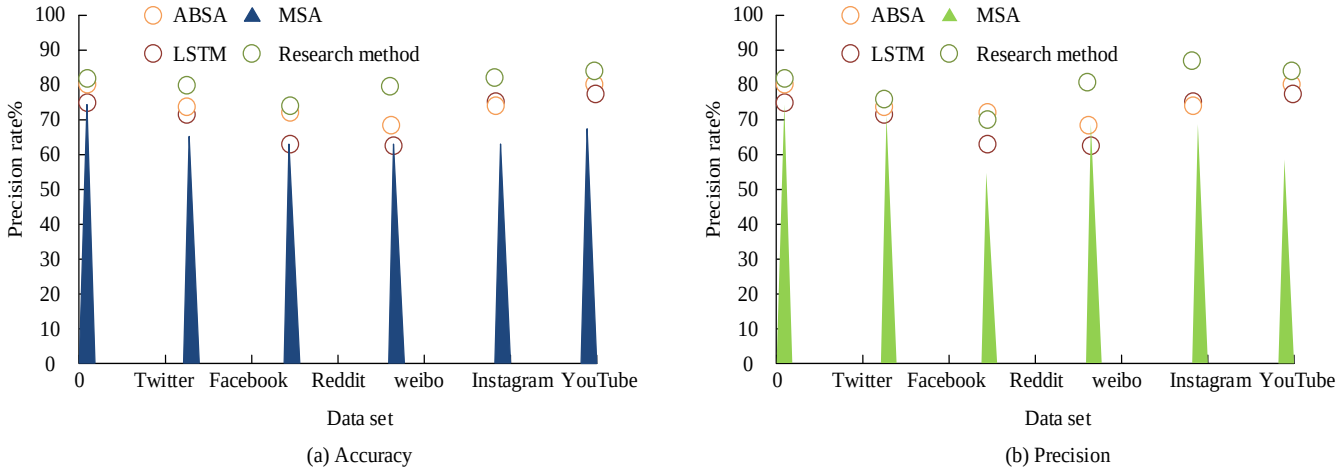


Fig. 6. Text visualization.



(a) Accuracy (b) Precision

Fig. 7. Comparison of precision and accuracy of four models.

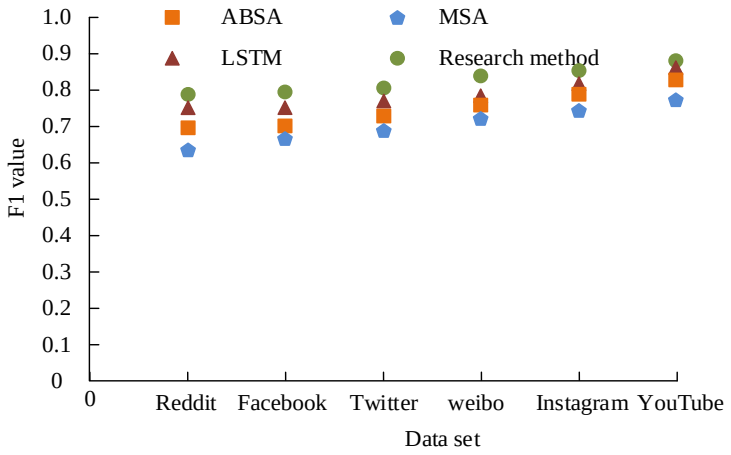


Fig. 8. Comparison of F1 values of four models on different datasets.

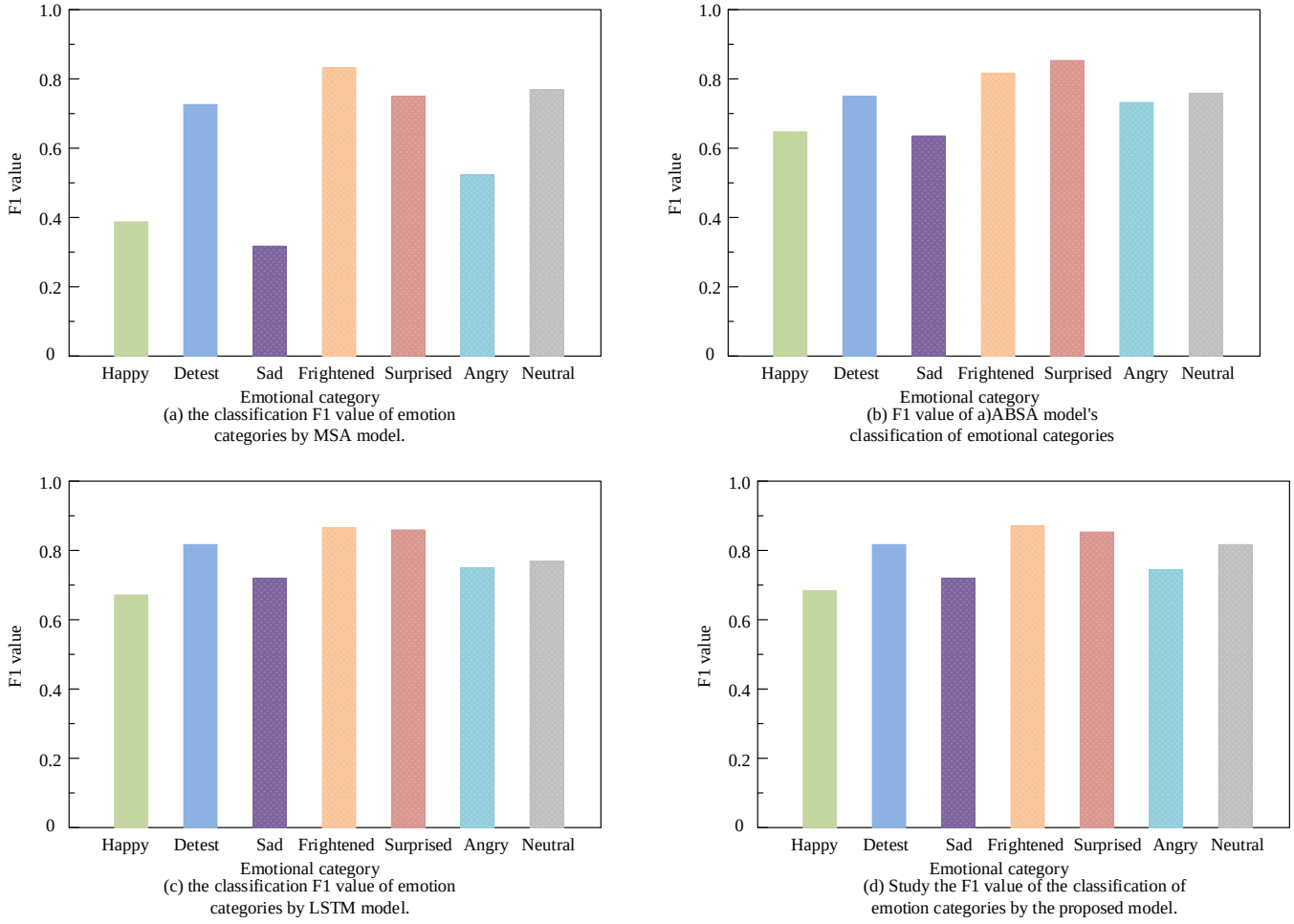


Fig. 9. Classification results of emotions using four models.

In Fig. 8, the F1 values of the four models on the social network dataset show similar performance to the precision shown in Fig. 6. Among them, the MSA model not only performs the worst in precision, but also has the lowest F1 value on the dataset, with an average F1 value of only 0.70. The F1 values of ABSA and LSTM are average, but LSTM performs slightly better than ABSA. The average F1 values for ABSA and LSTM are 0.74 and 0.78, respectively. The research model has relatively high precision and F1 values, with an average F1 value of 0.83 in the six datasets. Based on Fig. 8 and Fig. 7, the YouTube dataset was selected as a representative, and MSA, ABSA, LSTM, and research models were compared for emotion classification. The results are shown in Fig. 9.

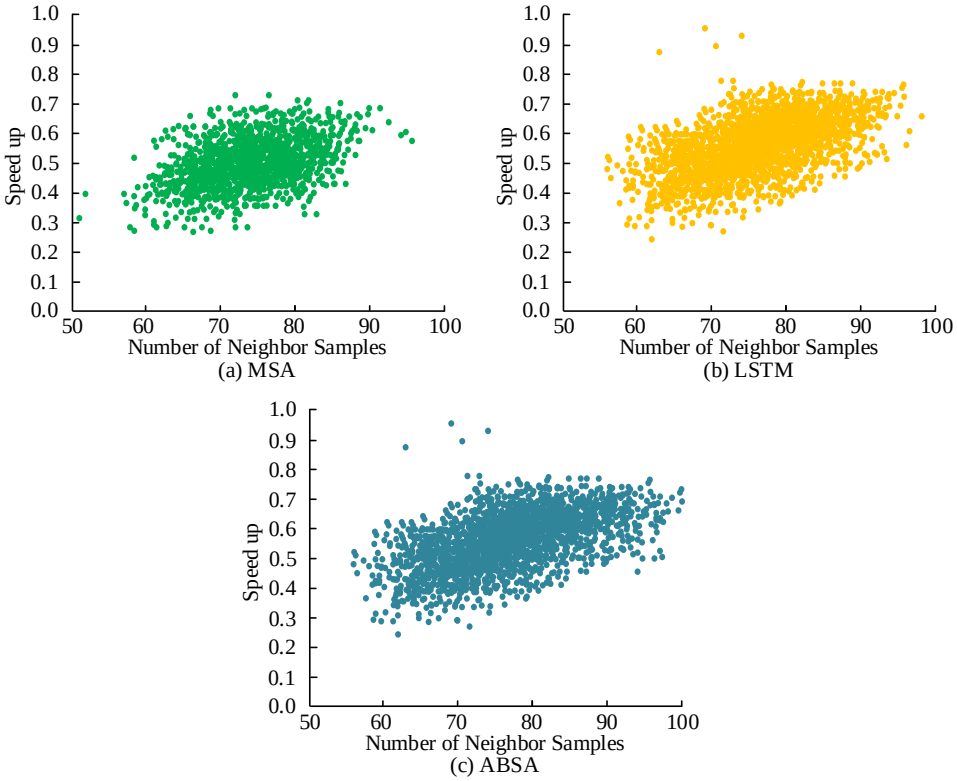


Fig. 10. Acceleration ratio results of different models.

In Fig. 9, compared to other models, the MSA model performs poorly in emotion recognition, indicating poor performance. This indicates that the early fusion and late fusion models have obvious shortcomings, which cannot balance the modeling of intra and inter modal features. Using the MSA model as a reference, there is a significant improvement in the recognition F1 values for happy, sad, and angry. Compared to others, the research model performs relatively well, especially in the F1 value of fear. The proposed attention based multi-level mixed fusion multi task TMSAM not only achieves the best experimental results in emotion classification, but also performs greater than the comparison method in emotion classification tasks such as detest, happy, sad, and surprise, fully verifying the effectiveness of the model. To measure the performance improvement of parallel computing compared to serial computing, Fig. 10 shows the comparative acceleration ratios of four models.

In Fig. 10, the acceleration ratios of MSA, ABSA, LSTM, and the study model are 0.73, 0.94, 1.17, and 1.61. The research model has the highest acceleration ratio and exhibits good parallel performance, which can save numerous training time when the training data is large.

V. Results and Discussion

The temporal multimodal sentiment analysis model based on multi task learning has shown significant performance advantages on social network datasets. The main reason is that the study has constructed a comprehensive framework based on convolutional networks, bidirectional gated recurrent units, and multi head self attention mechanism, targeting the characteristics of social network data. This fusion of multimodal data can fully utilize the correlation between different data modalities, improving the model's understanding ability for sentiment analysis tasks. The model proposed by the research institute achieved excellent results in accuracy and F1 values of 0.83 and 0.78, respectively. Compared to the system optimization research based on multimodal data fusion in study [23], our model performs better in sentiment classification tasks, which may be due to its more comprehensive modeling of inter modal features. Compared with the multimodal neural network semantic segmentation based on multi-scale RGB-T fusion in study [24], our model performs better in emotion recognition tasks. This may be because our model is more refined in the design of multimodal fusion and attention mechanisms. In addition, compared with the multi task learning model in study [25], our model performs better in both emotion and emotion classification tasks, possibly due to the use of a more suitable dataset and parameter settings for social network data, as well as a more effective modal fusion strategy. In summary, the time-series multimodal sentiment analysis model based on multitask learning proposed in the study can deeply understand the characteristics of data and achieve good sentiment analysis, thus having wide applicability in this field. However, the interpretability of the model has not been thoroughly analyzed in research, and further understanding of the data characteristics and user behavior of different social network platforms is needed to optimize model design. Future work should address these issues and enhance the interpretability and applicability of the model.

VI. Conclusion

In response to the problem of TMSAM in social networks, this study designed a sentiment analysis model based on MTL to fully utilize text and other modal information. It ensured the efficient performance of the model in social network

environments through detailed parameter tuning, and introduced attention mechanisms to enhance the model's ability to learn text embedding features. The results proved that for the MA of the model, MSA reached 0.69, ABSA was 0.78, and LSTM was 0.71, while the model proposed in the study was more advanced, reaching 0.83. In terms of average F1 value, the MSA was 0.70, ABSA was 0.74, LSTM was slightly higher than the former at 0.78, and the research model once again stood out with a high score of 0.83. The MA and F1 values of the research model were higher than those of other comparative models, highlighting the robustness and accuracy of the model. The acceleration ratios of MSA, ABSA, LSTM, and research models were 0.73, 0.94, 1.17, and 1.61, respectively. The acceleration ratio of the research model was the highest, and the comparative conclusions verified the advantages of multitasking learning and multi-modal fusion in improving parallel computing performance. This indicates that the research model exhibits excellent parallel performance in social network datasets, which can significantly save time when processing large-scale training data. In summary, the designed model is relatively reliable. TMSAM grounded on MTL is an attempt in social network sentiment analysis, providing a theoretical basis for effective solutions to complex sentiment classification problems.

However, there are still some problems and limitations to be solved. Among them, the interpretability of the model has not been deeply analyzed, which may affect its credibility and reliability in practical applications. In addition, the in-depth understanding of the data characteristics and user behavior of different social network platforms still requires more research to further optimize the model design. Future work can focus on solving these problems and exploring how to further improve the interpretability and applicability of the model to meet the actual needs of social network sentiment analysis tasks.

ACKNOWLEDGMENT

The research is supported by: The study was funded by the Shaanxi Normal University Graduate Student Pilot Talent Fund Project, ‘Research on the Influencing Factors of Language Learning Anxiety and Academic Achievement of International Students in China, (NO. LHRCTS23020).

REFERENCES

[1] Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 2023, 91: 424-444.

[2] Lyu Y, Schiopu I, Munteanu A. Multi-modal neural networks with multi-scale RGB-T fusion for semantic segmentation. Electronics Letters, 2020, 56(18): 920-923.

[3] Zhang J, Yin Z, Chen P, Nichele S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 2020, 59(1): 103-126.

[4] Pinto G, Carvalho J M, Barros F, Soares S. Multimodal emotion evaluation: A physiological model for cost-effective emotion classification. Sensors, 2020, 20(12): 3510.

[5] Zhao G, Luo Y, Chen Q, Qian X. Aspect-based sentiment analysis via multitask learning for online reviews. Knowledge-Based Systems, 2023, 264(5): 110326.1-110326.12.

[6] Rahmani S, Hosseini S, Zall R, Kangavari M, Kamran S, Hua W. Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects. Knowledge-Based Systems, 2023, 261(2): 110219.1-110219.16.

[7] Middya A, Nag B, Roy S. Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities. Knowledge-based systems, 2022, 244(3):108580.1-108580.14.

[8] Zhou H, Du J, Zhang Y, Wang Q, Liu Q, Lee C. Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29(7): 2617-2629.

[9] Zhang K, Zhu Y, Zhang W, Zhu Y. Cross-modal image sentiment analysis via deep correlation of textual semantic. Knowledge-Based Systems, 2021, 216(10): 106803.1-106803.12.

[10] Kumari R, Ashok N, Ghosal T, Ekbal A. Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. Information Processing & Management, 2021, 58(5):102631.

[11] Akhtar M S, Chauhan D S, Ekbal A. A deep multi-task contextual attention framework for multi-modal affect analysis. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(3): 1-27.

[12] Plaza-Del-Arco F M, Molina-González M D, Ureña-López L A, Martín-Valdivia M T. A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access, 2021, 9: 112478-112489.

[13] Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 10790-10797.

[14] Jiang H, Jiao R, Wang Z, Zhang T, Wu L. Construction and Analysis of Emotion Computing Model Based on LSTM. Complexity, 2021, 2021(4): 8897105-1-8897105-12.

[15] Zhang S, Yin C, Yin Z. Multimodal sentiment recognition with multi-task learning. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 7(1): 200-209.

[16] Bhosle K. and Musande V., Evaluation of Deep Learning CNN Model for Recognition of Devanagari Digit, Artif. Intell. Appl, 2023, 1(2): 114-118.

[17] Wang S, Zheng F, Zhao D. Research on Causal Network of High-dimensional Time Series with Insufficient Information. Journal of Chinese Computer Systems, 2023, 44(5): 981-990.

[18] Han Y, Liu M, Jing W. Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer. IEEE Access, 2020, 8: 21314-21325.

[19] Zhang X, Yu L, Tian S. BGAT: Aspect-based sentiment analysis based on bidirectional GRU and graph attention network. Journal of Intelligent & Fuzzy Systems, 2023, 44(2): 3115-3126.

[20] Younis M C, Abuhammad H. A hybrid fusion framework to multi-modal bio metric identification. Multimedia Tools and Applications, 2021, 80(17): 25799-25822.

[21] Zhou T, Fu H, Chen G, Shen J, Shao L. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. IEEE transactions on medical imaging, 2020, 39(9): 2772-2781.

[22] Cai G, Lyu G, Lin Y, Wen Y. Multi-level deep correlative networks for multi-modal sentiment analysis. Chinese Journal of Electronics, 2020, 29(6): 1025-1038.

[23] Gaw N, Yousefi S, Gahrooei M R. Multimodal data fusion for systems improvement: A review. IISE Transactions, 2022, 54(11): 1098-1116.

[24] Lyu Y, Schiopu I, Munteanu A. Multi-modal neural networks with multi-scale RGB-T fusion for semantic segmentation. Electronics Letters, 2020, 5(18): 920-923.

[25] Zhang Y, Wang J, Liu Y, Rong L, Zheng Q, Song D, Qin J. A Multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. Information Fusion, 2023, 93: 282-301.