Enhancing Emotion Prediction in Multimedia Content Through Multi-Task Learning

Wan Fan
Department of Culture and Communication, West Anhui University, Lu'an 237000, China

***Abstract*—This study presents a robust multimodal emotion analysis model aimed at improving emotion prediction in film and television communication. Addressing challenges in modal fusion and data association, the model integrates a Transformer-based framework with multi-task learning to capture emotional associations and temporal features across various modalities. It overcomes the limitations of single-modal labels by incorporating multi-task learning, and is tested on the Cmumosi dataset using both classification and regression tasks. The model achieves strong performance, with an average absolute error of 0.70, a Pearson correlation coefficient of 0.82, and an accuracy of 47.1% in a seven-class task. In a two-class task, it achieves an accuracy and F1 score of 88.4%. Predictions for specific video segments are highly consistent with actual labels, with predicted scores of 2.15 and 1.4. This research offers a new approach to multimodal emotion analysis, providing valuable insights for film and television content creation and setting the foundation for further advancements in this area.**

***Keywords*—Multi task learning; multimodal emotion analysis; timing; transformer; attention**

I. Introduction

With the rapid development of big data and artificial intelligence technologies, the application value of multimodal sentiment analysis in the field of film and television communication has become increasingly important. However, current studies are still lacking in multimodal information fusion, temporal feature modeling, and accurate capture of emotional expressions [1-2]. For example, many studies only rely on a single modality (e.g., textual or visual) for emotion recognition, failing to take full advantage of the synergy between different modalities, resulting in limited generalization ability of the models in complex scenes. Furthermore, extant methodologies demonstrate deficiencies in their capacity to effectively address the dynamic fluctuations in multimodal data, impeding the accurate capture of emotional nuances in film and television content. The paucity of labeled data further curtails the efficacy of model training [3-5]. Consequently, the efficient fusion of multimodal information, the enhancement of time-series modeling capability, and the mitigation of label scarcity problem are pivotal issues that must be addressed in the present context. To address the above challenges, this study proposes a temporal multimodal sentiment analysis model that combines Transformer structure and multi-task learning. The model uses Transformer to deep mine temporal features, and supplements the lack of unimodal labels with a self-supervised learning strategy to improve the accuracy of sentiment recognition. Meanwhile, the study adopts a multi-level feature fusion approach to enhance the complementarity of information between different modalities to improve the model's ability to capture dynamic changes in sentiment. In addition, a multi-task learning mechanism is introduced to enhance the adaptability of the model in different scenarios. The central objective of the research is to develop a model that can effectively integrate multimodal information and accurately analyze emotional changes. This model aims to address the issue of emotion recognition in film and television communication. The research has important theoretical and practical values. Theoretically, the research combines the transformer structure and multi-task learning to improve the existing multimodal sentiment analysis methods and provide a new technical framework for temporal feature modeling and multimodal fusion. Practically, the research results can be widely applied in the fields of film and television communication, social media sentiment analysis, online education, and mental health assessment. It can provide more accurate sentiment feedback for content creators, optimize user experience, and promote the further development of sentiment computing technology. The overall structure of the research includes seven sections: The initial parts summarizes the relevant research of multi-task learning and sentiment analysis. Section II, a temporal multimodal emotion analysis model based on multi task learning is proposed. Section III is to conduct experimental analysis on the proposed model. Section IV summarizes the experimental results, points out the shortcomings of the research, Discussion is given in Section V. Finally the paper is concluded in Section VI followed by future work in Section VII.

II. Related Works

With the enhancement of computing power and the surge in data volume, multi task learning gradually demonstrates its advantages in improving model efficiency and performance. Especially in the field of sentiment analysis, multi task learning was widely applied to understand and process various data modalities such as text, audio, and video [6-7]. Many scientists and research institutions have focused on developing advanced algorithms to more accurately identify and analyze human emotions. The following will introduce some related research by scientists and scholars. Yang et al. provided a new solution of a two-stage, multi-task multimodal emotional analyzing mechanism. It adopted a two-stage training strategy for full utilization. The experiment outcomes indicated that the proposed one outperformed the current model in most metrics on both datasets, proving its usefulness [8]. Barnes et al. provided a new solution of a multi task approach for sentimental analyzing solution. This model had hierarchical neural structures, and confirmed that using negation as a multi-task training was a useful way to upgrade the sentimental

analyzing performance, which had been proven on different datasets [9]. Yu et al. provided a new solution of a label generation solution to obtain unimodal supervision and emotional differentiation information, and conducted experiments on different datasets to jointly combine the modalities of multiple tasks. It demonstrated the reliability and stability of multimodal self-generated unimodal supervision [10]. Zhang et al. provided a new solution of a multi task learning multimodal solution. This model proposed cross modal attention and cross task attention. The former was designed to simulate multimodal feature fusion, while the other was designed to capture the interaction. The experiment outcomes indicated that this method had achieved good performance on the dataset assisting sentiment recognition [11].

Mamta et al. provided a new solution of a multimodal method for deep learning. They took the utilization of cross lingual word embeddings to map two languages to a shared semantic space on different languages. The experiment indicated that the proposed one had an accuracy of 0.70 on the movie review dataset, demonstrating good performance [12]. Gao et al. proposed a negative multi-task learning framework to alleviate gender bias. The experiment outcomes indicated that the proposed one could effectively alleviate bias while maintaining the utility of the model in sentiment analysis [13]. Yin et al. provided a new solution of a multi task joint sentiment analysis model. This model used joint training of multiple tasks to obtain local feature characteristics. The results indicated that the model proposed by the research was superior to state-of-the-art models [14]. Saha et al. provided a new solution of a multimodal speech act classification approach based on multi task learning. In addition, a network was combined to capture and effectively integrate shared semantic relationships across modalities. The experiment outcomes indicated that the proposed solution enhanced multimodal speech acts by benefiting from two secondary tasks compared to single modal and single task multimodal speech act classification [15].

In summary, although multi-task learning and sentiment analysis have been widely studied by numerous domestic and foreign experts and scholars, and have been successfully applied. However, shortcomings exists in the challenges of association mining and fusion, especially in the processing of multimodal data. Existing methods often struggle to effectively integrate interactive information between different modalities such as text, audio, and visual, thereby limiting the depth and accuracy of sentiment analysis. In response to this challenge, a temporal multimodal sentiment analysis model combining multi-task learning based on the Transformer framework is proposed. By combining multimodal fusion and temporal data analysis, this model provides a new method for complex sentiment analysis, which is of great positive significance for promoting sentiment analysis and improving the accuracy of sentiment recognition in film and television content.

III. Construction of the Temporal Multimodal Emotion Analysis Model based on Multitask Learning

The study first uses the Transformer framework to extract temporal features closely related to emotions. On this basis, a multi task learning temporal multimodal emotion analysis model is further proposed, aiming to solve the insufficient single modal label issue and improve the recognition and analysis ability of complex emotional states.

A. Time Series Multimodal Emotion Analysis Model Based on Transformer

In film and television communication, multimodal sentiment analysis refers to the use of different types of data, such as text, audio, and video analysis of emotions. Compared with single-modal methods, multimodal sentiment analysis can comprehensively utilize multiple sources of information and improve the accuracy of emotion recognition. To this end, a time-series multimodal sentiment analysis model based on Transformer is proposed, which uses linguistic guided Transformer (LGT) to explore emotional relationships between different modal data. Moreover, through the soft mapping (SM) module, emotion related features are mapped to higher dimensions, effectively integrating multimodal information. This method improves the accuracy of sentiment analysis and optimizes the processing of temporal data. Fig. 1 presents the architecture of this temporal multimodal emotion analysis model.

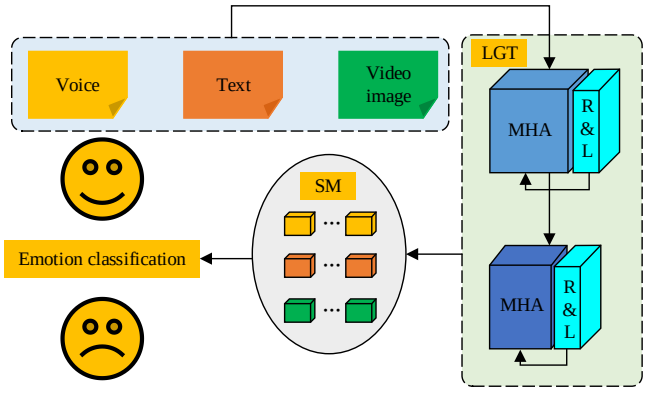


Fig. 1. Multimodal emotion analysis structure diagram.

In Fig. 1, the structure of this multimodal sentiment analysis model includes inputs in three modalities: speech, text, and video image. First, after feature extraction, each modal input is fed into the SM module, which maps the features of different modalities into a unified sentiment representation space. Then, the mapped features are fed into the multi-attention mechanism and the residual connection and normalization module to mine the correlations and affective features between different modalities. The language-guided framework further enhances affective feature extraction through the hierarchical multi-head attention (MHA) module. Finally, the model performs sentiment classification on the fused features and outputs the sentiment classification results. The traditional MHA mechanism is commonly used in machine translation, relying on parallel processing to accelerate the learning process. Its uniqueness lies in its ability to simultaneously process multiple data points without considering their temporal relationships [16-17]. The study applies this mechanism to multimodal emotion analysis to explore the mapping relationships between different modalities. Based on this logic, the LGT model is proposed, which utilizes MHA modules and forward neural networks (FNN) to explore emotional connections between different modalities. Although the modal data of this model

may not be synchronized in time, the model can process these data in parallel, effectively improving the efficiency of temporal data analysis. The architecture of LGT is shown in Fig. 2.

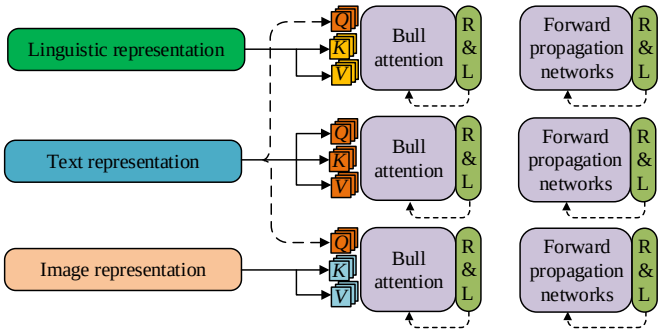


Fig. 2. Architecture diagram of LGT.

Fig. 2 shows a multimodal emotion analysis model dominated by text, supplemented by speech and image modalities. Each modal feature is independently inputted into the MHA mechanism. This setting allows emotional text data to guide speech and image data, thereby exploring the emotional connections between them. The text features are first decomposed into three vectors: query Q_l, key K_l, and value V_l, and these vectors are linearly transformed. Secondly, queries and key vectors are used to calculate attention scores. Finally, the attention score is combined with the value vector and the final result is calculated by weighted sum, as shown in Eq. (1).

$$Attention(Q_i, K_l, V_l) = soft\max((Q_l K_l^T)/\sqrt{d_k})V_l \quad (1)$$

In Eq. (1), d_k represents the dimension. This process is repeated multiple times, each representing an independent $head$. By combining the results of these heads, the final MHA output can be obtained, as shown in Eq. (2).

$$\begin{cases} head_i = Attention(Q_l W^Q, K_l W^K, V_l W^V) \\ F_{(l)} = MHA((Q_l, K_l, V_l)) = Concat(head_1, ..., head_h)W^O \end{cases} \quad (2)$$

After calculating the attention data, these results are immediately input into the FFN, aiming to explore the nonlinear connections between features and improve their expressive power. Eq. (3) provides the information of calculation process.

$$FFN = \mathrm{Re}lu(H'W^1 + b^1) + b^2 \quad (3)$$

The output of each layer of the LGT model is processed through residual connections and hierarchical normalization to ensure effective information transmission and stability, as shown in Eq. (4).

$$LayerNorm(x + Sublayer(x)) \quad (4)$$

In multimodal emotion analysis, when processing speech and image features, the MHA mechanism uses text modality as the source of the query vector. Meanwhile, keywords and truth vectors are derived from speech and image modalities. This method allows text features to introduce different representation spatial information [18], as shown in Eq. (5).

$$\begin{cases} F_{(a)} = MHA(Q_l, K_a, V_a) = Concat(head_1, ..., head_n)W^O \\ F_{(v)} = MHA(Q_l, K_v, V_v) = Concat(head_1, ..., head_n)W^O \end{cases} \quad (5)$$

In the next step of multimodal emotion analysis, the SM module is used to map the learned results of each modality to a new space, fuse them here, and then perform sentiment classification. Fig. 3 displays the schematic diagram of SM structure.

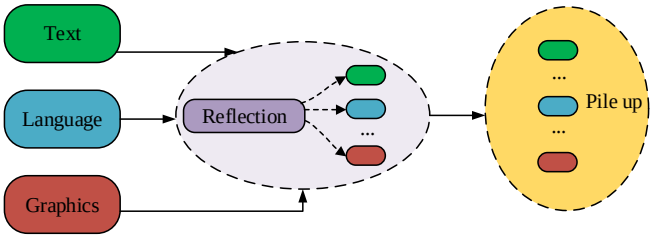


Fig. 3. SM structure diagram.

In Fig. 3, firstly, the input features of each modality, such as text, language, and graphics, enter the reflection module. It performs mapping processing on the input features and converts them into a unified representation format. It then stacks the mapped features by modality, laying the foundation for subsequent multimodal feature fusion and analysis. In the specific steps, the first step is to map the output from the FNN network to a higher dimensional space, and the calculation process is shown in Eq. (6).

$$NewMatrix = W_m M \quad (6)$$

In Eq. (6), W is the transformation matrix with a size of $2k \times k$, used to map the original matrix M to a higher dimension. Next, a vector set $\{v_i^p\}$ of $1 \times k$ size is used to perform soft attention calculation on each matrix in high-dimensional space, and these results are weighted and integrated into vector m_i to obtain the final calculation result of soft attention, as shown in Eq. (7).

$$\begin{aligned} p_i &= soft\max((v_i^p)^T (NewMatrix)) \\ SoftAttention_i(M) &= m_i = \sum_{j=0}^{N} (p_{ij} M_j) \end{aligned} \quad (7)$$

In the final stage of multimodal emotion analysis, the final output of SM is formed by stacking the results of the aforementioned computations. In addition, at the end of each step, a residual operation and hierarchical standardization processing are performed to ensure that the input of each round is integrated into the results of the previous round. Eq. (8) expresses the details.

$$\begin{cases} s = Stacking(\sum_{j=0}^{N} (m_j)) \\ M = LayerNorm(M + s) \end{cases} \quad (8)$$

In Eq. (8), M represents the output matrix, and s represents the independent output matrix obtained for each mode. In multimodal emotion analysis, it is necessary to accumulate the vectors corresponding to each modality in element order, and then perform classification prediction on this accumulated result according to Eq. (9).

$$y \sim p = W_p(LayerNorm(S_l + S_a + S_v)) \quad (9)$$

B. Time Series Multimodal Emotion Analysis Model Based on Multitask Learning

As a machine learning method, multi-task learning allows a model to perform several related tasks simultaneously in one training session, thereby improving the overall performance of the model. Transformer is a neural network structure for processing sequential data, such as text data, that can capture remote dependencies in the data, allowing the model to better understand contextual information. On the basis of a Transformer based temporal multimodal emotion analysis model, a self-supervised label generation module (SLGM) model with fusion multi-level correlation mining framework (MCMF) is proposed. The goal of MCMF is to explore the representation and correlation between multimodal data, improve existing models, and deeply explore emotional connections. Meanwhile, the study applies multi-task learning to multimodal emotion analysis to address the challenges of collaborative learning. By combining three single-modal tasks of text, speech, and image, collaborative learning is achieved through joint training using a multi-task learning framework. The goal of the SLGM model is to address the common problem of missing single modal labels in multimodal datasets. It generates single modal labels according to the mapping of modal representation and labels to meet the training requirements. Fig. 4 shows the structure of a temporal multimodal emotion analysis model based on multi task learning.

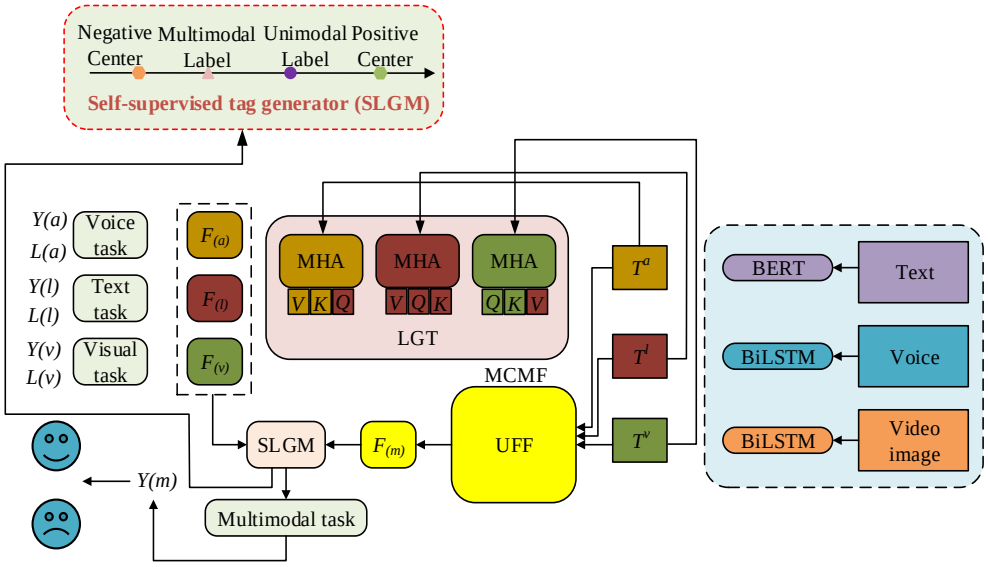


Fig. 4. Temporal multimodal emotion analysis supported by multi task learning.

In Fig. 4, the model mainly includes MCMF module and SLGM module. First, the model receives text, speech and visual data and extracts features through BERT, BiLSTM and other modules to solve the multimodal fusion problem in sentiment analysis. Second, SLGM generates unimodal and multimodal sentiment labels to compensate for the lack of label scarcity. Meanwhile, the unimodal features fusion (UFF) module in MCMF and LGT captures intermodal sentiment associations and fuses multimodal information through the MHA mechanism to improve the accuracy of sentiment prediction. Finally, the model realizes the comprehensive analysis of three-modal sentiment through multi-task learning. The research focuses on exploring the emotional connections between different multimodal representations from top to bottom. To discover the primary feature associations between these representations, a model agnostic fusion strategy is adopted in the study. Inspired by tensor fusion networks, UFF is used to overcome the limitations of traditional fusion methods, mapping single modal features to higher dimensions for fusion [19-20]. UFF fuses each single mode representation through triple Cartesian product, and captures the interaction between bimodal and trimodal through multi-level fusion, as shown in Fig. 5.

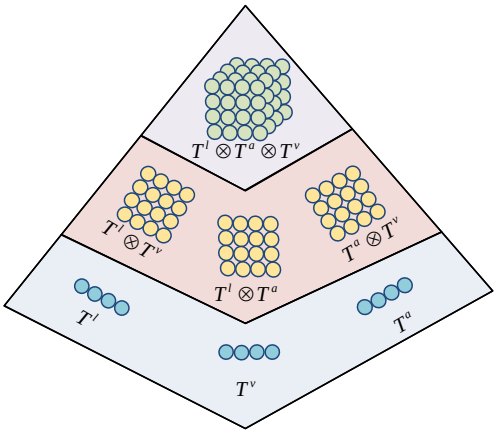


Fig. 5. Single mode feature mapping fusion UFF strategy.

In Fig. 5, the first step is to use T^l、T^a、T^v to capture the intrinsic characteristics of different modes. The second step focuses on exploring the interaction between the two modes, achieved through the calculation of $T^l \otimes T^a$, $T^l \otimes T^v$, and $T^a \otimes T^v$. Finally, the results of these stages are integrated to form a comprehensive fusion tensor, as shown in Eq. (10).

$$\begin{cases} \{(T^l, T^a, T^v) \mid T^l \in [{}_1^{T^l}], T^a \in [{}_1^{T^a}], T^v \in [{}_1^{T^v}]\} \\ F_{(m)} = [{}_1^{T^l}] \otimes [{}_1^{T^a}] \otimes [{}_1^{T^a}] \end{cases} \tag{10}$$

In Eq. (10), $\otimes$ represents the outer product operation, which is used to combine the features of multimodal m, text l, speech a, and visual v. By adding additional dimensions to the single modal tensor T^l、T^a、T^v and performing outer product operations, fused features are formed. Further application of hard parameter sharing mechanism in the framework to achieve sentiment analysis, where low-level networks share weights and high-level networks assign weights to specific tasks, as shown in Fig. 6.

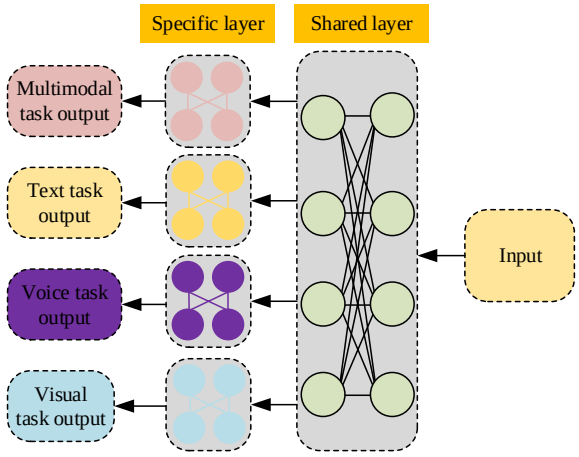


Fig. 6. Emotion analysis structure of multi task learning.

In Fig. 6, the study uses basic learning networks as shared layers, while independent prediction networks for each task are used as specific layers. Under this framework, a multimodal $F_{(m)}$ and three single modal tasks are set, with F^l, F^a, and F^v as inputs, respectively. Eq. (11) provides the definition of the dedicated layer.

$$\begin{cases} F_s^* = \mathrm{Re}\,LU(F_s W_s^{1T} + b_s^1) \\ y_s = F_s^* W_s^{2T} + b_s^2 \end{cases} \tag{11}$$

To adapt to multi-task learning frameworks, the SLGM model is proposed, which aims to generate single modal labels from multimodal data. SLGM is based on two core principles: the mapping relationship between modal representation and modal supervised values, and the proportional consistency of mapping relationships between different modalities. This method allows for the effective generation of labels required for single modal tasks, with specific calculations detailed in Eq. (12).

$$(F_m \# L_m) \infty (F_u \# L_u) \tag{12}$$

In Eq. (12), m represents multimodality and u represents covering $\{linguistic, acoustic, visual\}$ modes. F represents modal representation, L represents modal supervised value, $\#$ represents mapping relationship, and ∞ represents proportional relationship. These definitions indicate that modality and multimodal labels are highly correlated in terms of emotional polarity. However, in practical applications, there may be differences in the emotional polarity between single modal labels and multimodal labels. Among them, multimodal labels may be positive, but individual visual modalities may be marked as negative due to crying expressions. The strategy of SLGM to solve this problem is to divide modal representation as two categories abide by the emotional polarity, followed by calculating the center value of each category separately, and obtain the center representation of emotions from modalities, as shown in Eq. (13).

$$\begin{cases} C_p = \dfrac{\sum_{i=1}^{N} I(y(i) > 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) > 0)} \\ C_n = \dfrac{\sum_{i=1}^{N} I(y(i) < 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) < 0)} \end{cases} \tag{13}$$

In Eq. (13), $I(\cdot)$ is the symbol of the indicator function, N is the symbol of the number, and F represents the modal representation. Next, SLGM applies the Bartcharia distance coefficient for measuring the sample and its corresponding category center. The specific calculation is detailed in Eq.(14).

$$\begin{cases} S_p = \sum_{j=1}^{K} \sqrt{F(j) C_p(j)} \\ S_n = \sum_{j=1}^{K} \sqrt{F(j) C_n(j)} \end{cases} \tag{14}$$

In Eq. (14), K is the symbol of the total elements number in the modal representation. SLGM uses proportion and difference to reflect the mapping relationship, so the original Eq. (12) has been updated and expressed as Eq. (15).

$$\begin{cases} S_m / L_m = S_u / L_u, S_m - L_m = S_u - L_u \\ L_u = L_m + \dfrac{S_u - S_m}{2} * \dfrac{L_m + S_m}{S_m} \end{cases} \tag{15}$$

In Eq. (15), S_m and S_u represent the deviation degrees of multimodal and single-mode, respectively, while L_m and L_u refer to the markings of multimodal and single-mode. To address the issue of result fluctuations caused by iterations in the label generation process, SLGM has implemented a mechanism for dynamically updating single modal labels. The specific calculation method is shown in Eq. (16).

$$y_u^{(i)} = \frac{1}{2} * \frac{i-1}{i+1} * y_u^{(i-2)} + \frac{1}{2} * \frac{i-1}{i+1} * y_u^{(i-1)} + \frac{2}{i+1} * y_u^i, i \geq 3 \quad (16)$$

In summary, starting from 3rd iteration, each iteration i's outcome is associated with the first two rounds of $(i-1)$ and $(i-2)$, ensuring that as the number of iterations increases, the generated labels gradually stabilize.

IV. Test of Multimodal Emotion Analysis Model in Film and Television Communication

The research experiment first sets two datasets and their experimental parameters to conduct the experiment. Secondly, the results on these two datasets are compared and ablation experiments are performed to analyze the contributions of each component. Subsequently, performance comparisons are made with other models. Finally, the study conducts an analysis of the effectiveness of LGT and selects some film and television clips for in-depth analysis, demonstrating the emotional analysis ability of the proposed model.

A. Test of a Transformer Based Time-Series Multimodal Emotion Analysis Model

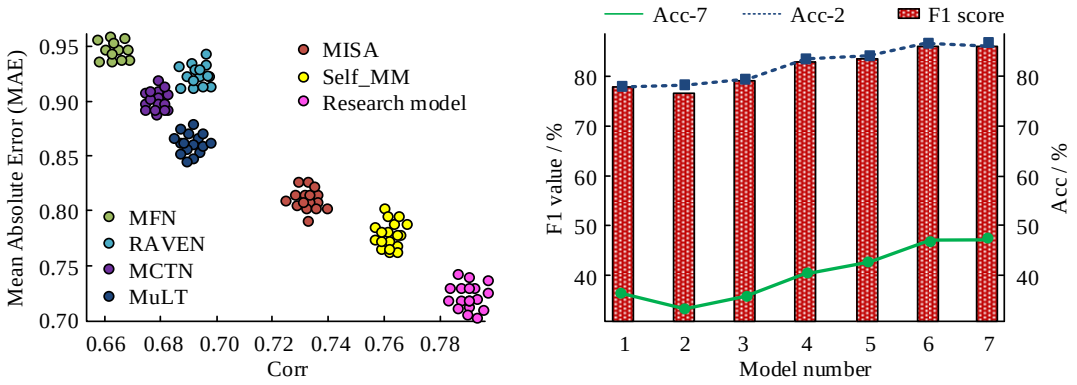
To validate the temporal multimodal emotion analysis model, two datasets, Cmumosi and Cmumosei, are selected for the study. The CMUMOSI and CMUMOSEI datasets contain 2,200 video clips and 24,000 YouTube comment clips, respectively. These cover a variety of data formats, including text, audio, and video modes. Bidirectional encoder representations from Transformers are used for feature extraction of text data, while Mel-Frequency Cepstral Coefficients are used for feature transformation of audio data. Video is sampled at 2 frames per second, scaled to 224x224 pixels, and visual features are extracted using a convolutional neural network. To ensure modal synchronization, the window alignment method is used to process multimodal data. In addition, to solve the problem of unbalanced emotion category, the study further applied a small amount of data enhancement strategy to improve the model's low-frequency emotion recognition ability.

Since the final model consists of MCMF and SLGM, MCMF includes UFF and LGT, and LGT consists of SM, MHA and FNN, ablation experiments are set up to test the benchmark performance of each module, and the results are shown in Table I. The LGT model based on Model 1, which consists of SM, MHA, and FNN, can extract basic emotional features. However, its performance is mediocre due to the lack of modal fusion and label generation. After UFF module is added to LGT in Model 2, the correlation between modes is enhanced and various indicators are improved. SLGM module is added to LGT in Model 3 to make the model capable of generating single mode labels, which effectively improves the accuracy of sentiment analysis. Model 4 is an MCMF module composed of UFF and LGT. Through the high-dimensional fusion of UFF and the emotional feature analysis of LGT, the performance of the model is close to the final model, and the synergistic effect of the two is verified. Model 5 is the final model. Combined with SLGM, the accuracy of multimodal emotion recognition is further improved. Moreover, the accuracy rate, recall rate and F1 score are all the best, which are 0.92, 0.93, and 0.92, respectively.

TABLE I. Ablation Test Results

Network structure	Accuracy value	Recall value	F1 score
SM+MHA+FNN (Model 1: LGT base model)	0.83	0.82	0.81
UFF+SM+MHA+FNN (Model 2: Add UFF module)	0.86	0.85	0.85
LGT+SLGM (Model 3: LGT combined with SLGM module)	0.88	0.87	0.87
UFF+LGT (Model 4: MCMF module)	0.89	0.89	0.88
UFF+LGT+SLGM (Model 5: MCMF+SLGM complete model)	0.92	0.93	0.92

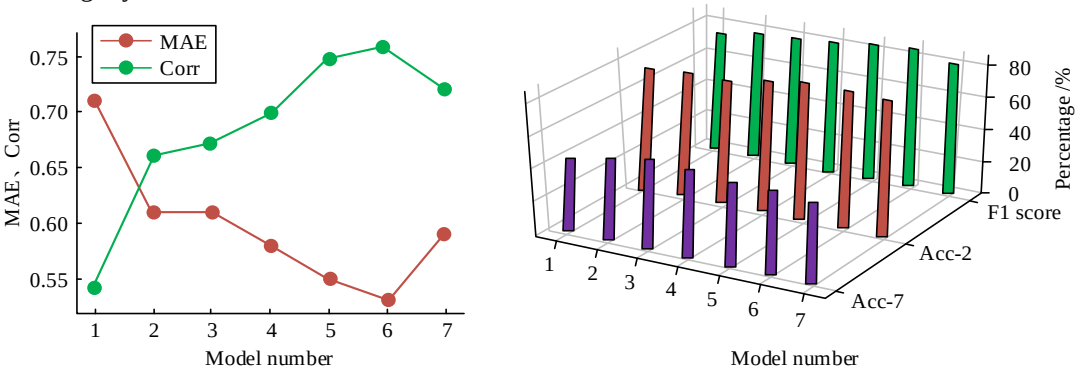
The testing is divided into regression and classification tasks, which hires Pearson correlation (Corr) together with mean absolute error (MAE) as indicators. The classification task is evaluated using seven class accuracy (Acc-7), two class accuracy (Acc-2), and F1 score (F1). The proposed one is planned to be compared with multiple existing models on the Cmumosi dataset, as shown in Fig. 7. The compared models include memory fusion network (MFN), recurrent attention variation embedding network (RAVEN), multimodal cyclic translation network (MCTN), multicurrent Transformer (MulT) Modality Invariant and Specific Representations for multimodal emotion analysis (MISA) and self-supervised multi task learning for multimodal emotion analysis (Self MM). These models are denoted as Model 1 to Model 6, respectively. In Fig. 7(a) and Fig. 7(b), the multi-task learning time-series multimodal sentiment analysis model proposed in the research outperforms the existing mainstream models in several key metrics. In the regression task, the MAE of the proposed model is 0.70, which is significantly lower than the 0.87 of MulT and the 0.78 of MISA. It indicates that the research model has a smaller error in sentiment prediction accuracy. Meanwhile, the Corr of the research model reaches 0.82, which is higher than the 0.70 of MulT and 0.76 of MISA. It indicates that the research model is more robust in modeling emotional features and can more accurately reflect the emotional associations between different modalities. In the categorization task, the Acc-7 of the research model is 47.1%, which is significantly better than the 42.3% of MISA and the 46.8% of Self-MM, indicating that the proposed model is able to discriminate different categories of emotions more accurately. In addition, the research model achieves 88.4% in both Acc-2 and F1 values, surpassing the 86.1% of Self-MM and the 83.1% of MulT. This result further validates the advantages of the multi-task learning framework in emotion recognition, which can effectively improve the ability to capture complex emotional features.



(a) MAE and Corr for Cmumosi (b) Acc-7, Acc-2, and F1 score for Cmumosi

Fig. 7. Comparison on the cmumosi dataset.

The next comparison on the Cmumosei dataset is shown in Fig. 8. In terms of the regression task, the MAE of the research model drops to 0.59, which is lower than both 0.61 for MCTN and 0.71 for MFN. It implies that its prediction of sentiment trends is more accurate. Meanwhile, the Corr of the research model reaches 0.72, which is significantly better than the 0.54 of MFN) and 0.67 of MCTN. It implies that the research model is able to extract multimodal features more stably and establish reasonable sentiment mapping relationships among different modalities. In the classification task, the research model's Acc-7 is 49.6%, which is slightly lower than MulT's 51.8% and Self-MM's 52.2%, but reaches 82.2% on Acc-2, which is higher than MCTN's 79.3% and MFN's 77.9%. Meanwhile, its F1 score is 82.2%, which showed a strong sentiment categorization ability. In summary, although the research model approaches the optimal model in some indicators, its overall performance exceeds that of most mainstream methods, particularly in the regression task. The research model demonstrates a more stable and robust sentiment prediction ability, ensuring accurate sentiment analysis of film and television communication content.



(a) MAE and Corr for Cmumosei (b) Acc-7, Acc-2, and F1 score for Cmumosei

Fig. 8. Different models are compared on the cmumosei dataset.

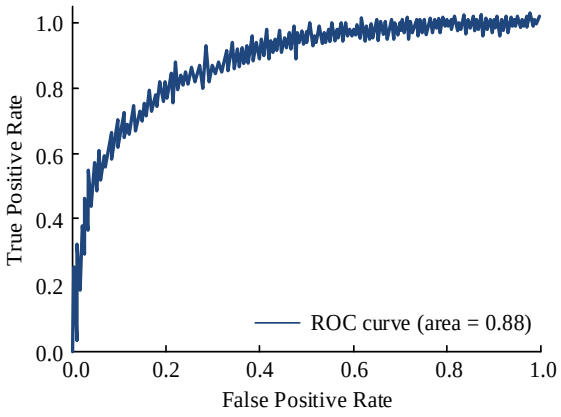


Fig. 9. Receiver operating characteristic curve.

Next, how the research model performing is validated by randomly selecting over 4000 samples from the test set for binary classification results testing. Through assessing the predicted labels with the real sample labels together, the true and false positives are calculated. Based on this data, the operating characteristic curves of the subjects are plotted, and the outcomes are given in Fig. 9. The model proposed in the study has a strong ability to distinguish between two types of emotional polarity, reflecting its advantage in accurately identifying and predicting emotional tendencies in film and television content.

The research model combines two mechanisms, LGT and SM, respectively, for modal interaction and mapping results to high-dimensional space for effective fusion. The study conducted ablation analysis on the Cmumosi dataset, as shown in Fig. 10. Both of these structures have a positive impact on

sentiment classification. The experiment is divided into 1-4 scenarios: no LGT and SM, only SM, only LGT, and both LGT+SM are used. In Fig. 10(a) and Fig. 10(b), the model performance decreases significantly with the removal of LGT and SM, indicating the importance of these two modules in sentiment modeling. When using only SM, although it improves some tasks, the overall performance is still limited by the lack of intermodal interaction capability. Using only LGT, although it improves modal interaction, it lacks the support of the soft-mapping mechanism, resulting in a decrease in the ability to fuse high-dimensional features. The complete model (LGT+SM) performs optimally and achieves the best results in all metrics, confirming the key role of LGT in enhancing intermodal affective interactions, while the SM module enhances the feature expression capability through high-dimensional mapping. The results of the ablation experiments illustrate that the combination of the two can maximize the emotion recognition ability of the model, making it more applicable to complex film and television communication scenarios.

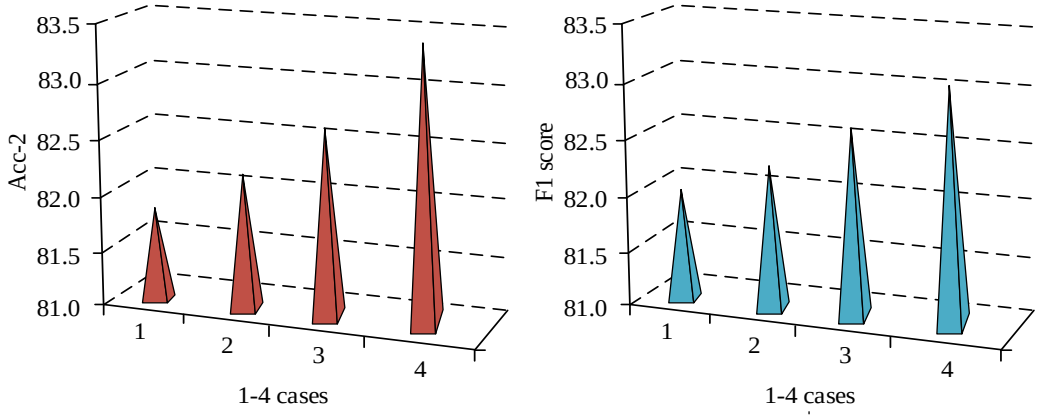


Fig. 10. Cmumosi dataset ablation analysis.

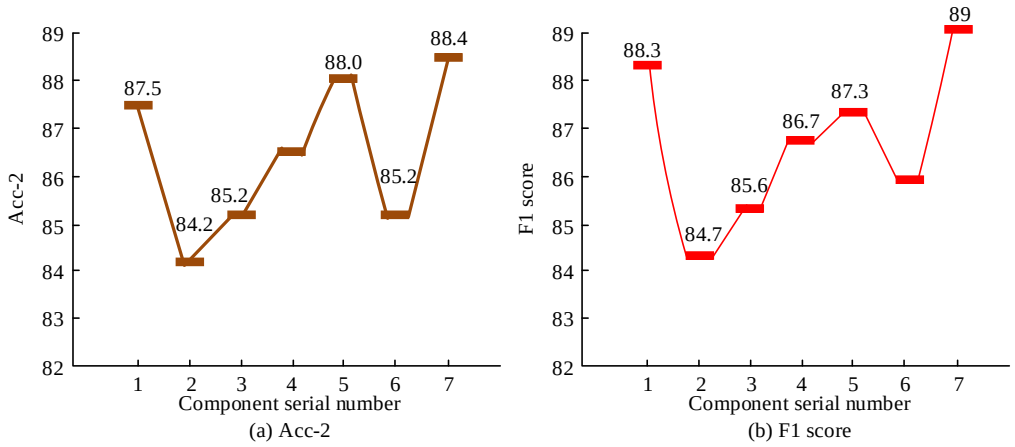
B. *Test of a Time-Series Multimodal Emotion Analysis Supported by Multitask Learning*

To compare the performance of different models, classification and regression evaluation indicators are used in the study, as shown in Table II. The research model outperforms the comparison model in all indicators: Acc-2 reaches 88.5%, which is 2.4% higher than Self_MM and 1.9% higher than Zhang Y et al.'s model. It indicates that the sentiment polarity determination is more accurate. Acc-7 reaches 47.2%, which outperforms MISA, Self_MM, and Yu W et al.'s model. It indicates that it has a stronger generalization ability in the task of complex sentiment classification. MAE is the lowest at 0.70, with better error control than Barnes J et al. (0.74) and Yang B et al. (0.76), which is more stable in prediction. Corr is the highest at 0.82, which is better able to accurately capture the trend of emotion change compared to Zhang Y et al.'s model at 0.79 and Self_MM's 0.80. Overall, the research model performs best on all indicators, validating the effectiveness of multi-task learning and temporal multimodal fusion methods to provide more accurate prediction results in complex film and television sentiment analysis tasks.

This study conducted ablation tests on the Cmumosi dataset to accomplish the effects assessment of various components of the model. The experiment set up seven components: UFF, LGT, SLGM, UFF+LGT, UFF+SLGM, LGT+SLGM, and UFF+LGT+SLGM, numbered 1 to 7, respectively. Fig. 11 gives the information that feature concatenation is used when UFF is missing, single modal representation is directly input when LGT is missing, and multimodal labels are used for training when SLGM is missing. The results show that UFF is the key to performance improvement, especially with the combination of UFF and SLGM achieving the highest Acc-2 and F1 scores of 88.0% and 89.0%, respectively. The combination of UFF+LGT and LGT+SLGM did not perform as expected, with 86.5% and 85.2%, respectively, due to the lack of single modal labels or basic correlations between modalities.

TABLE II. DIFFERENT MODELS ON THE CMUMOSI DATASET

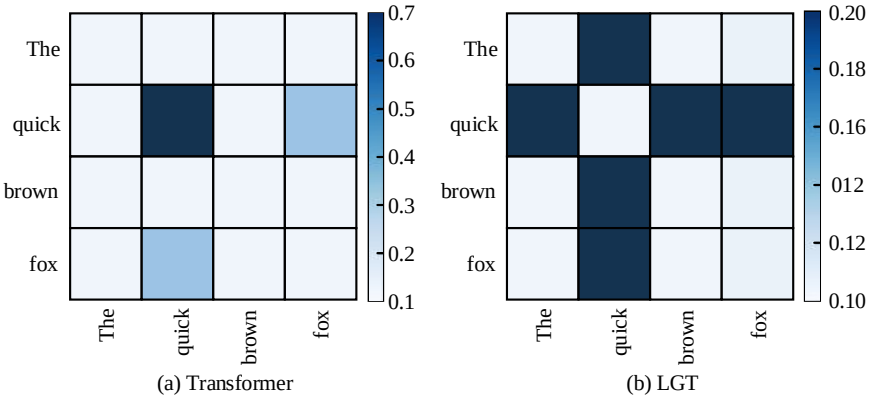
Type	Acc-2	Acc-7	MAE	Corr
MFN	77.9	36.3	0.95	0.66
RAVER	78.1	33.2	0.92	0.69
MCTN	79.3	35.6	0.91	0.68
MulT	83.1	40.2	0.87	0.7
MISA	83.4	42.3	0.78	0.76
Self_MM	86.1	46.8	0.71	0.8
The model of Yang et al.	85.6	45.59	0.74	0.8
The model of Barnes et al.	86.05	45.17	0.74	0.8
The model of Yu et al.	84.96	45.67	0.73	0.79
The model of Zhang et al.	84.42	46.01	0.72	0.79
Research model	88.5	47.2	0.7	0.82



(a) Acc-2 (b) F1 score

Fig. 11. Ablation analysis of different component structures.

For confirming the LGT model effectiveness, a test sentence "The quick brown fox" is used on the Cmumosi dataset, and the performance of LGT and Transformer in attention mechanism is compared. As shown in Fig. 12, the attention scores are represented by color depth, where a darker color is the symbol for higher weights and a lighter color is the symbol for lower weights. Comparing Fig. 12(a) and Fig. 12(b), LGT is more accurate in assigning word weights. For example, in the MHA mechanism of Transformer, "quick" pays significant attention to "fox", while the distribution of LGT is more uniform, allowing for a more balanced focus on the structure of the entire sentence. This demonstrates the advantage of LGT in finely adjusting attention weights, especially in capturing key vocabulary. It further demonstrates its effectiveness in processing language tasks, especially in understanding sentence structures.



(a) Transformer (b) LGT

Fig. 12. LGT attention matrix result graph.

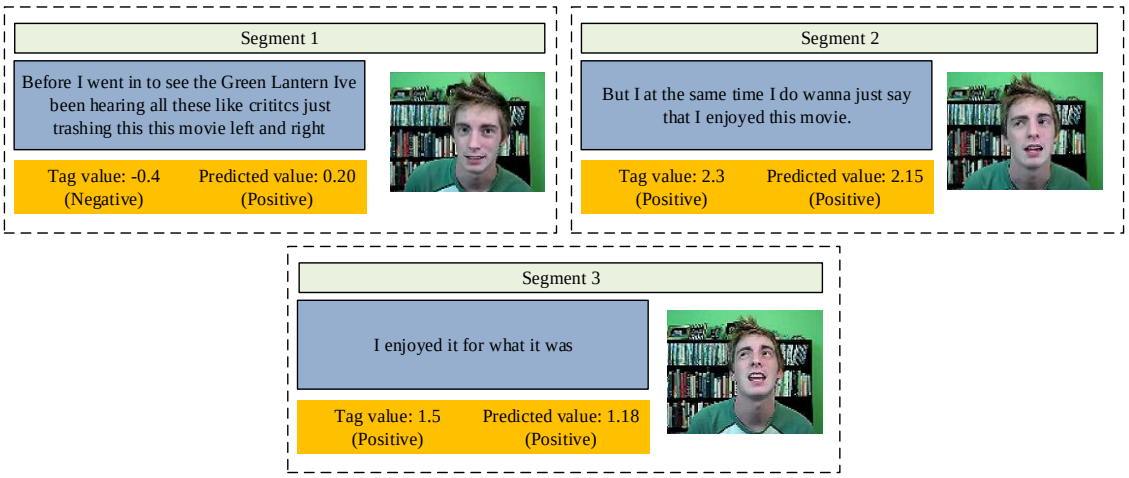


Fig. 13. Cmumosi dataset film and video emotion analysis results.

The study selects specific film and television video clips from the Cmumosi dataset to accomplish the evaluation of the the proposed one's analytical ability. Three video clips are selected for detailed testing in the experiment, with text content and keyframe images shown in Fig. 13, supplemented by experimental labels and prediction results. The observation can tell that the predictions of the second and third segments are highly consistent with the actual labels, demonstrating the accuracy of the model in sentiment analysis, with predicted scores of 2.15 and 1.18, respectively. However, there is a deviation in the results of the first segment, with a predicted value of only 0.20. This difference is attributed to the fact that the image at the beginning of the video can mislead the model's judgment. The combination of the smiling image of the man in the first clip with neutral text leads the model to lean towards positive predictions. In summary, this experiment reveals the complexity of the model in interpreting emotions and demonstrates its advantages in integrating visual and textual cues.

TABLE III. TEST RESULTS OF IEMOCAP AND MELD DATASETS

Type	Acc-2	Acc-7		F1-Score	MAE	Corr
MFN	0.78	0.65		0.72	0.9	0.65
RAVER	0.79	0.66		0.73	0.88	0.67
MCTN	0.8	0.68		0.75	0.85	0.69
MulT	0.83	0.7		0.78	0.82	0.72
MISA	0.85	0.72		0.8	0.78	0.74
Self_MM	0.86	0.73		0.82	0.76	0.76
The model of Yang et al.	0.84	0.71		0.81	0.78	0.76
The model of Barnes et al.	0.83	0.71		0.82	0.77	0.75
The model of Yu et al.	0.83	0.73		0.82	0.77	0.75
The model of Zhang et al.	0.84	0.71		0.83	0.77	0.73
Research model	0.88	0.75		0.85	0.73	0.78

To further verify the applicability of the model in different areas of emotion analysis, the study introduces additional datasets of IEMOCAP and MELD to test the performance of the model in conversational emotion recognition and multimodal emotion dialog, respectively. The IEMOCAP dataset contains multimodal dialog emotion labels, which can be used to test the emotion recognition effect of the model in the dialog scene. The MELD dataset provides multimodal emotion data of multi-round dialogues, which can be used to verify the emotion recognition ability of the model in complex scenes. In Table III, the proposed model outperforms the other models in all indicators on both IEMOCAP and MELD datasets. The models proposed by Yang B et al. Barnes J et al. Yu W et al. and Zhang Y et al. shows strong competitiveness in Acc-2, Acc-7, and F1-Score, but compared to the proposed model, they still fall slightly short of the proposed model, especially in Acc-2 and F1-Score. In particular, in terms of Acc-2 and F1-Score, the proposed model reaches 0.88 and 0.85, while the highest Acc-2 of the other models is only 0.84, and the highest F1-Score is 0.83. It suggests that the proposed model has more advantageous in terms of sentiment categorization accuracy. In addition, the MAE value of the proposed model is 0.73, which is more stable and exhibits a lower error rate compared to the 0.77-0.78 of the other models. It proves its reliability in fine-grained emotion recognition tasks. Overall, the proposed model of the study demonstrated better performance in the sentiment analysis task, further validating its application value in the field of film and television communication.

V. DISCUSSION

To improve the accuracy of multimodal sentiment analysis in film and television communication, this study presented a time-series model based on a Transformer structure with MCMF and SLGM modules. Compared with the two-stage multi-task learning model proposed by Yang B et al. [8], which improved feature classification, this approach addressed limitations when multimodal labels were sparse. By incorporating the SLGM module to generate single-modal labels, this model not only enriched the diversity of emotional labels, but also significantly improved the accuracy of complex emotion recognition. Results on IEMOCAP and MELD datasets showed F1 scores up to 0.85. In terms of efficiency, the model reduced the MAE on the CMUMOSEI dataset by 15%, demonstrating the effectiveness of the multi-task learning framework in sentiment analysis. Compared to Zhang Y et al.'s model [11], which used cross-modal attention and multi-task learning for accuracy, this model improved label stability and captures subtle emotions, achieving 47.1% accuracy in seven-category tasks and 88.4% accuracy in binary tasks. This improvement was due to the SLGM module, which ensured efficient emotion prediction even with sparse single-modal labels. In practical applications on social media platforms, the model could be utilized to analyze the emotional trends of users' video content, thereby providing real-time emotional feedback to content creators and platform operators for the purpose of optimizing content push strategies. In online education, the model could help identify students' emotional responses during video lessons so that teachers could adjust their teaching methods. In healthcare, the model helped assess mental health and monitor emotional states by analyzing video and voice recordings of patients.

For broader applications, the model could be extended to recognize multi-character interactions and nuanced emotions. Such scenes in film and television often involve multiple, complex emotions, where the recognition of subtle emotional states was crucial. Future studies could include these scenarios to verify the adaptability of the model. For real-time applications, the model showed high computational efficiency in binary emotion tasks, indicating potential for real-time deployment. Further optimization with a lightweight Transformer structure and a multi-tasking strategy could support real-time video analysis and meet the needs of efficient media analysis.

In summary, there are three main arguments of the study: First, SLGM improves the acquisition quality of unimodal sentiment labels and still maintains a stable sentiment recognition ability in multimodal datasets without complete annotation. Second, UFF allows different modal features to be fused more efficiently and improves the synergy between multimodal data, thus enhancing the robustness and generalization ability of the model. Third, MCMF enhances the

modeling capability of complex sentiment features by jointly optimizing LGT and SLGM, while improving the sentiment classification accuracy. These results show that the proposed model of the study not only outperforms existing methods on several benchmark datasets, but also makes a significant breakthrough in the accuracy and stability of multimodal data fusion and sentiment recognition. It lays a foundation for the application of sentiment analysis technology in more complex multimodal data applications, which has positive implications for improving the communication effects of movies and television and optimizing user experience.

VI. Conclusion

A temporal multimodal emotion analysis model supported by multi task learning was proposed to address the challenges of sentiment association mining and fusion in multimodal data. This model utilized a Transformer based framework to deeply explore temporal features related to emotions, and compensated for the shortcomings of single modal labels, enhancing the accuracy of emotion prediction. The experimental results showed that, specifically, in the binary sentiment analysis task on the CMUMOSEI dataset, the average processing time of the model was about 0.05 s per frame of data, enabling it to perform sentiment prediction in near real-time conditions. Furthermore, the ablation experiments demonstrated that the structural optimization of the UFF and LGT modules markedly reduced the computational overhead of the model and enhanced its computational speed by approximately 30% in comparison to traditional multimodal sentiment analysis models. It substantiated the superior performance of the proposed model over several existing models. These results demonstrated their clear advantages in accurately recognizing affective tendencies. To further validate the effectiveness of LGT, a comparative analysis was conducted on the differences in attention mechanisms between LGT and traditional Transformers. It was found that LGT was more precise in finely adjusting word weights, which was crucial for capturing the emotional meaning of sentences. In addition, by selecting specific film and television video segments for testing, the predictions of the second and third segments were highly consistent with the actual labels, with predicted scores of 2.15 and 1.4, respectively. It was proved that the research model can highly match the actual labels in most cases.

VII. Future Work

Despite the merits of the current model's superior emotion recognition performance as demonstrated in this study, there are notable deficiencies. First, the Transformer structure's application in temporal modeling incurs substantial computational complexity when dealing with video data of considerable duration. This may impede the efficiency of practical applications. Second, the multimodal fusion process primarily relies on global feature matching, neglecting the potential benefits of local information in sentiment analysis. This may result in a reduction in accuracy for certain nuanced sentiment recognition tasks. In addition, the research model still has room for improvement in the recognition of extreme emotion categories (e.g., intense anger or extreme happiness), and more fine-grained feature extraction methods can be explored in the future to improve the model's emotion classification ability. To address the aforementioned limitations, future research can be optimized in the following ways. First, a more efficient attention mechanism should be introduced to reduce computational complexity and improve the applicability of the model to long time series data. Second, local feature modeling should be combined to improve the model's sensitivity to subtle sentiment changes. Third, an adaptive multi-task learning strategy should be introduced to enhance the model's generalization ability in different application scenarios. Furthermore, at the level of the data, the diversity of the training data can be expanded in the future to encompass a more extensive range of movie and television styles and cultural backgrounds, thereby enhancing the robustness of the model.

In summary, this study improves the performance of multimodal sentiment analysis through an innovative model design, while pointing out the existing limitations and proposing appropriate improvement directions. The research results not only provide an effective solution for sentiment computation in the field of film and television communication, but also lay the foundation for the further development of multimodal sentiment analysis.

Acknowledgment

The research is supported by The Youth Project of Philosophy and Social Sciences Planning in Anhui Province in 2022: Research on the Influence of Social Short Video on the National Identity of Small Town Youth in Anhui Province (Project Number: AHSKQ2022D021).

References

[1] Singh A, Saha S, Hasanuzzaman M, Dey K. Multitask learning for complaint identification and sentiment analysis. Cognitive Computation, 2022, 1(1): 1-16.

[2] Zhang T, Gong X, Chen C L P. BMT-Net: Broad multitask transformer network for sentiment analysis. IEEE transactions on cybernetics, 2021, 52(7): 6232-6243.

[3] Mao R, Li X. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(15): 13534-13542.

[4] Bie Y, Yang Y. A multitask multiview neural network for end-to-end aspect-based sentiment analysis. Big Data Mining and Analytics, 2021, 4(3): 195-207.

[5] Chen F, Yang Z, Huang Y. A multi-task learning framework for end-to-end aspect sentiment triplet extraction. Neurocomputing, 2022, 479(3): 12-21.

[6] Zhao M, Yang J, Qu L. A multi-task learning model with graph convolutional networks for aspect term extraction and polarity classification. Applied Intelligence, 2023, 53(6): 6585-6603.

[7] Purohit J, Dave R. Leveraging Deep Learning Techniques to Obtain Efficacious Segmentation Results. Archives of Advanced Engineering Science, 2023, 1(1): 11-26.

[8] Yang B, Wu L, Zhu J, Shao B, Lin X, Liu T Y. Multimodal sentiment analysis with two-phase multi-task learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30(1): 2015-2024.

[9] Barnes J, Velldal E, Øvrelid L. Improving sentiment analysis with multi-task learning of negation. Natural Language Engineering, 2021, 27(2): 249-269.

[10] 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.

[11] Zhang Y, Rong L, Li X, Chen R. Multimodal sentiment and emotion joint analysis with a deep attentive multi-task learning model. European Conference on Information Retrieval. Cham: Springer International Publishing, 2022, 1(1): 518-532.

[12] Mamta, Ekbal A, Bhattacharyya P. Exploring Multi-lingual, Multi-task, and Adversarial Learning for Low-resource Sentiment Analysis. Transactions on Asian and Low-Resource Language Information Processing, 2022, 21(5): 1-19.

[13] Gao L, Zhan H, Sheng V S. Mitigate gender bias using negative multi-task learning. Neural Processing Letters, 2023, 55(8): 11131-11146.

[14] Yin C, Chen Y, Zuo W. Multi-task deep neural networks for joint sarcasm detection and sentiment analysis. Pattern Recognition and Image Analysis, 2021, 31(8): 103-108.

[15] Saha T, Upadhyaya A, Saha S, Bhattacharyya P. A multitask multimodal ensemble model for sentiment-and emotion-aided tweet act classification. IEEE Transactions on Computational Social Systems, 2021, 9(2): 508-517.

[16] Tan Y Y, Chow C O, Kanesan J, Chuah J H, Lim Y. Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning. Wireless personal communications, 2023, 129(3): 2213-2237.

[17] 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(1): 282-301.

[18] Maity K, Kumar A, Saha S. A Multitask Multimodal Framework for Sentiment and Emotion-Aided Cyberbullying Detection. IEEE Internet Computing, 2022, 26(4): 68-78.

[19] Abdullah T, Ahmet A. Deep learning in sentiment analysis: Recent architectures. ACM Computing Surveys, 2022, 55(8): 1-37.

[20] Hande A, Hegde S U, Chakravarthi B R. Multi-task learning in under-resourced Dravidian languages. Journal of Data, Information and Management, 2022, 4(2): 137-165.