TGMoE: A Text Guided Mixture-of-Experts Model for Multimodal Sentiment Analysis

Xueliang Zhao[1], Mingyang Wang[2]*, Yingchun Tan[3], Xianjie Wang[4]
College of Computer and Control Engineering, Northeast Forestry University, Harbin, China[1,2,3]
Harbin Institute of Technology, Harbin, China[4]

***Abstract*—Multimodal sentiment analysis seeks to determine the sentiment polarity of targets by integrating diverse data types, including text, visual, and audio modalities. However, during the process of multimodal data fusion, existing methods often fail to adequately analyze the sentimental relationships between different modalities and overlook the varying contributions of different modalities to sentiment analysis results. To address this issue, we propose a Text Guided Mixture-of-Experts (TGMoE) Model for Multimodal Sentiment Analysis. Based on the varying contributions of different modalities to sentiment analysis, this model introduces a text guided cross-modal attention mechanism that fuses text separately with visual and audio modalities, leveraging attention to capture interactions between these modalities and effectively enrich the text modality with supplementary information from the visual and audio data. Additionally, by employing a sparsely gated mixture of expert layers, the TGMoE model constructs multiple expert networks to simultaneously learn sentiment information, enhancing the nonlinear representation capability of multimodal features. This approach makes multimodal features more distinguishable concerning sentiment, thereby improving the accuracy of sentiment polarity judgments. The experimental results on the publicly available multimodal sentiment analysis datasets CMU-MOSI and CMU-MOSEI show that the TGMoE model outperforms most existing multimodal sentiment analysis models and can effectively improve the performance of sentiment analysis.**

***Keywords*—Multimodal fusion; sentiment analysis; cross modal; mixture of experts**

I. Introduction

With the rapid growth of text data such as social media and online comments, sentiment analysis has become an increasingly important research field. The goal of sentiment analysis tasks is to classify the sentiment information contained in raw data into different sentiment polarities such as positive, negative, or neutral. However, in many practical scenarios, sentiment data often not only contains textual information but also includes multimodal data such as images, videos, audio, etc. Compared to unimodal data lacking diversity, these multimodal data can provide more information for sentiment analysis, and the complementarity of this information can enhance the accuracy of sentiment analysis.

Existing multimodal sentiment analysis methods include Tensor-based fusion [1], which directly connects feature tensors from different modalities for analysis. However, this method generates very large feature tensors, requiring a lot of storage space and computational resources, and does not consider the interaction of information between different modalities. To address these issues, researchers have developed other deep learning-based fusion methods. Huddar et al. [2] proposed multi-level feature optimization, extracting feature tensors from multiple modalities and using LSTM to extract contextual information between adjacent utterances at multiple levels. However, this method did not examine the correlation of different modal information with the sentiment analysis results, making it unable to fully understand the target sentiment comprehensively and accurately. Tsai et al. [3] proposed a multimodal routing method that dynamically adjusts the relative weights between input samples and output representations by exploring the correlation between modalities and identifying the relative importance of single-modal and cross-modal features.

However, although previous approaches have made progress in multimodal fusion, they often fail to adequately account for the varying contributions of different modalities to sentiment information, overlooking the importance of sentimental information from different modalities. In the field of sentiment analysis, while audio and visual modalities indeed contain crucial sentimental information, the distribution of sentiment information across modalities is unevenly distributed. Neglecting the differences in contributions of different modalities to sentiment analysis may result in multimodal fusion representations lacking crucial sentiment information from key modalities, thus reducing the accuracy of sentiment analysis [4].

To address the above issues, this paper proposes a text guided mixture-of-experts model for multimodal sentiment analysis. The model aims to better capture the differences in sentiment information between different modalities, obtaining more targeted sentiment features. TGMoE leverages pre-trained models for feature extraction from three modalities. It integrates visual and audio modality information into the text modality through a text guided cross-modal fusion mechanism to obtain multimodal fusion features. Subsequently, for the sentiment prediction task, multiple highly specialized experts are simultaneously trained by a trainable gating network to selectively handle sentiment features. This approach delves deeper into uncovering potential connections among modal data, thereby enhancing the accuracy of sentiment prediction in the model.The contributions of TGMoE model can be summarized as follows:

- Proposing a text guided cross-modal Transformer network that integrates sentiment information from visual and audio modalities into the text modality through a text guided attention mechanism.
- TGMoE uses a sparsely gated mixture-of-experts mechanism to selectively process multi-modal fusion

features, enhancing the model's ability to learn and represent complex emotional information.

- Extensive experimental results on two benchmark datasets demonstrate that the proposed TGMoE outperforms several existing methods in multimodal sentiment analysis tasks.

II. Related Work

With the advent of the information age, we have access to a large amount of multimodal data (videos, audio, and text), providing a more abundant source of features for sentiment analysis tasks. Accurate and rapid analysis of human emotions can offer better services for daily work and life. Multimodal sentiment analysis aims to understand the sentiment of the target in video data (including text, audio, and visual modalities), uncovering deep sentimental information in each modality to reduce bias in single-modal sentimental information. Learning how to capture interaction information within and between modalities and effectively integrate multimodal information is a key challenge faced by multimodal sentiment analysis tasks [5].

To address this issue, researchers have proposed various multimodal fusion methods for modeling. With the advancement of deep learning, model-based fusion has received more attention. Zadeh et al. [1] introduced the Tensor Fusion Network, which computes the Cartesian product of three modalities and concatenates the resulting tensor along a certain dimension. The concatenated tensor is then fed into a deep neural network for sentiment classification. Building on this, Liu et al. [6] decomposed the weights of the fusion tensor into a set of low-rank factors to improve efficiency, making the computational complexity linearly related to the number of modalities. Hou et al. [7] recursively integrated local correlations into global correlations through multilinear fusion. Mai et al. [8] adopted a divide-and-conquer approach by partitioning multimodal features into blocks, applying tensor fusion to each block to capture local interaction information, and then combining local information to obtain global multimodal interaction information.

Model-based fusion methods can effectively preserve sentimental information within modalities but struggle to consider contextual relationships between modalities. Graph neural networks, due to their excellent structural learning capabilities, are widely applied in multimodal sentiment analysis tasks. Yang et al. [9] proposed a modal-temporal attention graph for unaligned multimodal data, where each sub-feature of each modality in the sequential data is treated as a node. They construct modality-type edges between different modalities and temporal-type edges within the same modality. By applying a pruning algorithm, they fuse and align asynchronous distributed multimodal sequential data. Hu et al. [10] constructed a fully connected heterogeneous graph for conversational data, considering each modality of each utterance as a node. They connect each node to nodes representing the same utterance but from different modalities and to nodes representing the same modality from the same conversation. Different aggregation mechanisms for various types of edges are designed to learn multimodal dynamics in the graph network.

To further explore sentimental information within and across modalities, researchers have started integrating attention mechanisms into multimodal sentiment analysis. For example, Wang et al. [11] proposed a recurrent attentional variable embedding network that combines attention mechanisms to investigate facial and speech features. They learn displacement information generated during the vocabulary representation process for multimodal representation. Building upon this, Rahman et al. [12] incorporated multimodal representations into Transformer models, using visual and audio features to learn representation shifts and apply them to text modality for sentiment analysis. While these methods have partially addressed the issue of insufficient fusion between modalities, they often treat each modality equally, overlooking the varying contributions of different modalities to the final sentiment analysis results.

Therefore, this paper addresses the issue of disparate contributions between different modalities by enhancing the role of the text modality in sentiment analysis. It utilizes a mixture of experts to further extract sentiment information from multimodal features, enhancing nonlinear representation capabilities and obtaining more abstract fusion features of sentiment information.

III. Method

A. Overall Model Architecture

Multimodal sentiment analysis employs three modalities - audio (X_a), visual (X_v), and text (X_t) - from the same video segment to determine the sentiment polarity of the target. The proposed TGMoE model also aims to effectively integrate information from the three modalities to enhance the effectiveness of sentiment analysis. Fig. 1 illustrates the overall architecture of the TGMoE model. The model framework consists of three parts: the feature extraction module, the text guided cross-modal feature fusion module, and the sparsely gated mixture of experts module.

Feature Extraction Module: For each modality, appropriate pre-trained models are used to gain incipient modality features.

Text Guided Cross-Modal Feature Fusion Module: Utilizing cross-modal attention to capture the interaction information between audio, visual, and text features, the module adds this information to the text features. This encourages the text features to incorporate information from other modalities, providing high-quality fused features for sentiment prediction.

Sparsely Gated Mixture-of-Experts Module: Training multiple experts to handle multimodal features with different sentimental biases, to deeply explore the potential connections between data and improve the accuracy of sentiment prediction.

B. Feature Extraction Module

To better extract features of single modality data, different feature extraction methods are adopted for different modalities. For the text modality, the language pre-trained model SimCSE [13] is utilized as the feature extractor for text discourse. The hidden state output from the last layer is taken as the feature vector for the text modality.

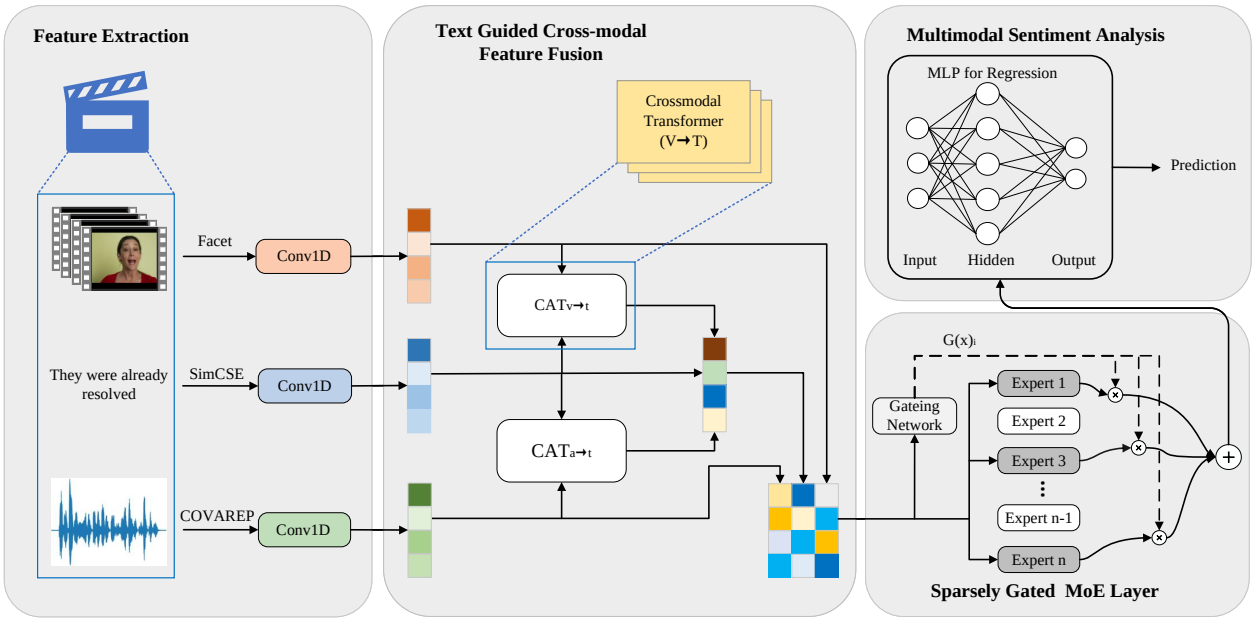

Fig. 1. The overall structure of the TGMoE model.

$$X_t' = \text{SimCSE}(X_t; \theta_t) \in \mathbb{R}^{s_t \times d_t} \tag{1}$$

Where X_t' represents the result of text modality feature extraction and θ_t is the SimCSE model's parameter. s_t means the sequence length and d_t is the feature dimension of text modalities.

For the audio modality, audio features are extracted using the COVEREP [14] acoustic framework. These features include pitch, volume, Mel-Frequency Cepstral Coefficients, and more, denoted as X_a. For the visual modality, visual features are extracted using Facet. These features include facial action units, facial landmarks, head pose, and other features, denoted as X_v. The features for audio and visual modalities can be obtained through the CMU-Multimodal SDK.

To achieve better fusion in the upcoming work, 1D temporal convolution is used to unify the feature dimensions of the three modalities while ensuring that each element of the input sequence has sufficient awareness of its neighboring elements. The features of the three modalities are fed into a 1D temporal convolutional layer:

$$f_m = \text{Conv1D}(X_m', k_m) \in \mathbb{R}^{s_m \times d_m}, m \in \{t, v, a\} \tag{2}$$

where k_m is the size of the convolutional kernel. f_m is output of 1D temporal convolutional layer.

C. Text Guided Cross-Modal Feature Fusion Module

In the traditional Transformer model, changing the positions of the input sequence does not alter the final output. To enable the model to capture the sequential information of the input sequence, positional embeddings (PE) are added to the representation of each modality based on the practice outlined in Transformer [15]:

$$H_m = f_m + \text{PE}_m \in \mathbb{R}^{s_m \times d_m}, m \in \{t, v, a\} \tag{3}$$

where PE_m means the PE of each modal, H_m represents the feature vector after each modal adds PE.

The text modality is the most basic and intuitive form of reflecting the speaker's sentiment, containing more sentiment-related information compared to video and audio modalities. Therefore, based on the idea of MulT [16], this paper presents a text guided cross-modal fusion module, which utilizes cross-modal attention mechanisms to calculate the attention weights between the text modality and the audio-visual modalities. This promotes the reception of information from the other two modalities by the text modality, potentially integrating emotion-related features from the audio and visual modalities into the text features for better encoding of emotional information across all three modalities. Additionally, considering the characteristics of the sentiment analysis task, the dominant role of the text modality in the feature fusion process is reinforced to incorporate emotional information from different modalities. The text-guided cross-modal feature fusion is illustrated in Fig. 2. Each cross-modal Transformer consists of L layers of cross-modal attention.

$$\text{CAT}_{v \to t} = \text{Softmax}(\frac{W_Q H_t \times W_K^T H_v}{\sqrt{d_k}}) W_V H_v \tag{4}$$

$$\text{CAT}_{v \to t} = \text{LN}(\text{CAT}_{v \to t}) \tag{5}$$

Where $\text{CAT}_{v \to t}$ indicates the visual modality transmission of information to the text modality. Softmax(·)represents a normalized exponential function, it can compress a K-dimensional vector z containing arbitrary real numbers into another K-dimensional vector $\sigma(z)$ such that the range of

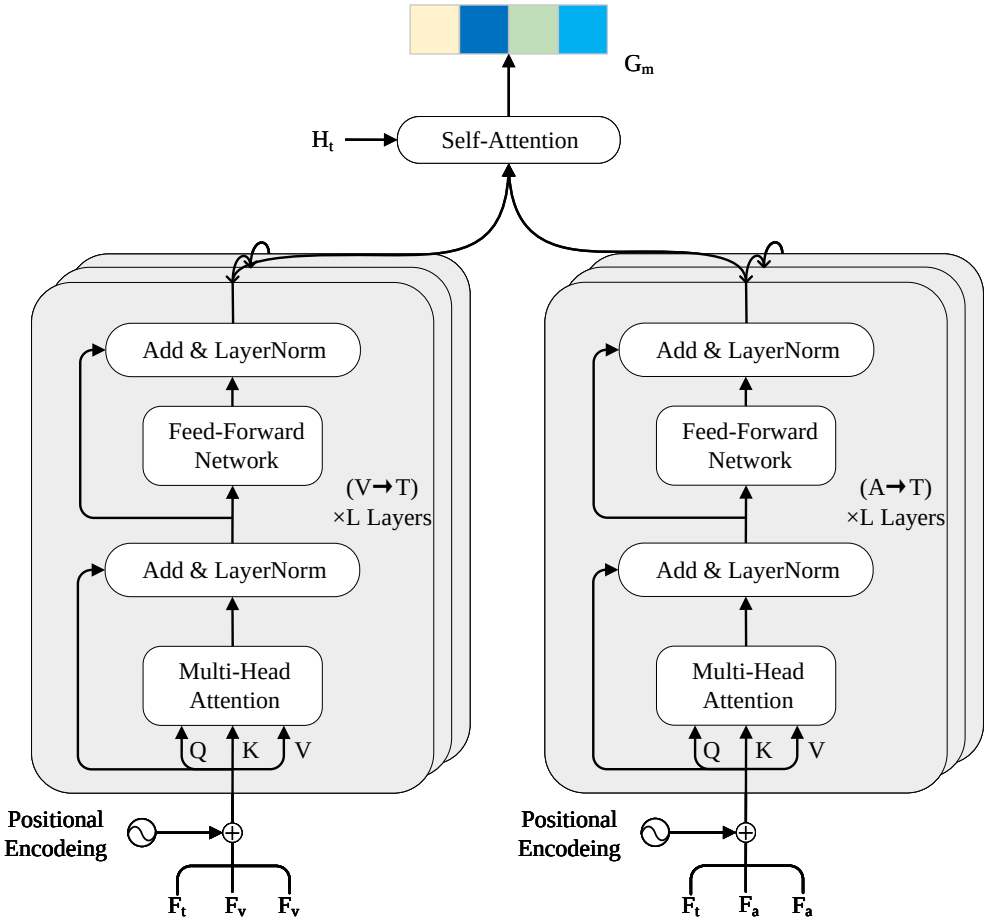

Fig. 2. Text guided cross-modal feature fusion module.

each element is between (0,1), and the sum of all elements is 1. The attention mechanism uses numbers between 0 and 1 to represent the importance of different data blocks. Where $Q = W_Q H_t, K = W_K^T H_v, V = W_V H_v$.

As each layer of the network can capture different abstract levels of input data, we stack L layers of cross-modal Transformers to learn hierarchical sentimental information within the features. Subsequently, cross-modal attention is further computed between the audio modality and the text modality, as exemplified by $(A \rightarrow T)$ in Fig. 2. The two fusion results are then concatenated with the text features and passed through self-attention blocks to capture the interactive information between the text features and the fusion features. Finally, the fusion features are combined with the features from the three modalities through convolutional layers to generate the ultimate fused feature representation. Adapting from the low-level features is beneficial for the model to retain the original information of each modality.

$$Z_m = H_t + \text{CAT}_{v \rightarrow t} + \text{CAT}_{a \rightarrow t} \tag{6}$$

$$G_m = \text{Self-Attention}(Z_m) \tag{7}$$

$$F_m = [G_m, H_a, H_v, H_t] \tag{8}$$

During the text guided cross-modal feature fusion process, the text modes constantly update their sequences through external information from multi-head cross-modal attention. By taking advantage of the fact that the text contains more sentiment-related information [17], the sentimental information of the text features is strengthened, and the sentimental information of the vision and audio modes is fully integrated into the text modes to obtain multi-modal features containing more sentimental information.

D. Sparsely Gated Mixture-of-Experts Module

During the process of text guided cross-modal feature fusion, sentiment may be expressed differently across different modalities. For example, in the sentence "You look beautiful today," the sentiment conveyed in the text modality is positive, but if accompanied by a pouting expression, the sentiment

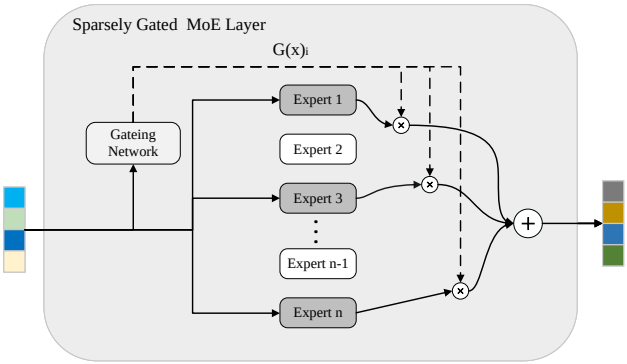

Fig. 3. Sparsely gated mixture-of-experts layer.

in the visual modality becomes negative. There are certain differences in the distribution of sentiment across different samples of data, so sentimental information in multi-modal features can be enhanced through a feed-forward network (FFN). However, for traditional deep learning models, activating the entire model for each sample when the training data is large can lead to significant spatial and time costs. To address this issue, the TGMoE model introduces a neural network component: the mixture of experts (MoE). By using gating mechanisms, the number of experts involved in the work can be effectively controlled, thus compressing the model's computational costs.

MoE is a special type of feed-forward network. In this structure, each model unit is referred to as an expert, and there is a gating network to select a combination of experts, combining the weight of each model as the final output. The difference from training data individually in traditional FFNs is that the mixture of experts' networks can enhance the non-linear representation capability of multi-modal features by allowing multiple experts to learn simultaneously. This enhances the distinctiveness of multi-modal features in terms of sentiments, thereby improving the classification accuracy of data samples.

The MoE layer in the TGMoE model consists of several experts and a trainable gating network. Each expert is an independent FFN that learns similar or different features from each other. The gating network learns parameters to select a sparse combination of numerous experts to process each input. The output of the gating network is a sparse n-dimensional vector, which is used to weigh the selected combination of experts. Each expert has the same architecture but distinct parameters. The structure of the MoE layer is illustrated in Fig. 3.

For a given input x, we define $G(x)$ as the output of the gating network; $E_i(x)$ is the output of the i-th expert network. Therefore, the output of the Sparsely Gated Mixture-of-Experts module is:

$$y = \sum_{i=1}^{n} G(x)_i E_i(x) \tag{9}$$

When $G(x)_i = 0$, the model does not need to compute $E_i(x)$. Therefore, although the model includes numerous neural networks, only a small number of neural networks will be utilized for each sample, significantly reducing computational complexity and time cost.

Sparsely Gated Network: The input x is multiplied by a trainable weight matrix W_g, then the initial architecture of the gating network is completed by applying the Softmax function.

$$G(x) = \text{Softmax}(x \cdot W_g) \tag{10}$$

x represents the input of sparsely gated network, and W_g represents a randomly generated matrix. Constructed in this way, the gating network outputs a non-sparse vector. Therefore, to ensure the sparsity of the gated output, the top k values of the gated output are retained. In addition, adjustable Gaussian noise is introduced to ensure that each neural network receives roughly the same amount of training data. The amount of noise for each gate is controlled by another adjustable parameter matrix W_{noise}.

$$\text{KeepTopK}(v_i, k) = \begin{cases} v_i, & v_i \in \text{TopK}(k) \\ -\infty, & v_i \notin \text{TopK}(k) \end{cases} \tag{11}$$

$$H(x) = x \cdot W_g + \text{LN}(x \cdot W_{noise}) \tag{12}$$

$$G(x) = \text{Softmax}(\text{KeepTopK}(H(x), k)) \tag{13}$$

where $\text{LN}(\cdot)$ represents data standardization, $v_i \in \text{TopK}(k)$ means that v_i belongs to the top K elements.

The MoE layer is placed after the text guided cross-modal feature fusion module. After passing through the text guided cross-modal fusion module layer, each multimodal fusion feature will invoke the MoE once, thereby selecting different combinations of experts to enhance the sentimental information in the multimodal representation. For the multi-modal feature F'_m, multiple expert networks are simultaneously trained through the MoE layer to deeply explore the potential connections between data, as described below:

$$\begin{aligned} z = \sum_{i=1}^{n} \text{Softmax}(\text{KeepTopK}((F_m \cdot W_g \\ + \text{LN}(F_m \cdot W_{noise})), k)) E_i(F_m) \end{aligned} \tag{14}$$

E. Model Prediction and Loss Function

The vector z output from the MoE layer is fed into a Multilayer Perceptron (MLP) for sentiment prediction. The MLP consists of three linear layers. The TGMoE model uses MAE as the basis for computing the loss function for the entire task.

$$Y'_m = \text{MLP}(Z_m; \theta_m), m \in \{t, v, a\} \tag{15}$$

$$\text{Loss}_m = \frac{1}{N} \sum_{i=1}^{N} (|\text{pred}^i - y^i|) \tag{16}$$

where N represents the number of training samples, and y represents the true labels of the multimodal data.

IV. Experimental Results

A. Datasets and Evaluation Metrics

1) Datasets: This paper evaluates the performance of the TGMoE model using two publicly available datasets.

The CMU-MOSI dataset [18] is a commonly used dataset in the field of multimodal sentiment analysis. This dataset consists of 93 videos from YouTube where the reviews of movies are discussed. The dataset is divided into 2,199 subjective discourse-level video segments. Each segment has a real-valued sentiment score in the range of [-3, +3] to express the intensity of the sentiment polarity of the characters.

The CMU-MOSEI dataset [19] is an extension of the CMU-MOSI dataset, with a larger number of utterances, more diverse samples, speakers, and topics. This dataset contains 23,453 video segments with sentiment-labeled tags from 5,000 videos. Each video segment in the MOSI and MOSEI datasets contains a sentiment score in the range of [-3, 3], where a higher value indicates a stronger positive sentiment polarity.

2) Evaluation metrics: When considering sentiment analysis on the CMU-MOSI and CMU-MOSEI datasets as a regression task, the predictive performance of the models is measured using Mean Absolute Error (MAE) and Pearson correlation (Corr). When viewed as a classification task, evaluation methods include seven-class accuracy (Acc-7), binary accuracy (Acc-2), and F1 score. Here, Acc-7 represents the accuracy of predicting values falling into seven intervals within [-3, +3], while Acc-2 and F1 represent the accuracy and F1 score of the binary classification task, respectively.

B. Experimental Settings

We developed the TGMoE model in the PyTorch framework, using Mean Absolute Error (MAE) as the loss function, Adam as the optimizer, and PyCharm as the integrated development environment. All experiments in this study were conducted on an RTX 4090 GPU, and multiple validations and analyses were performed to obtain the best set of hyperparameters.

We set the batch size to 32, epochs to 50, learning rate to 1e-3, and sequence length to 50 for all three modalities. The text feature dimension is 768, the visual feature dimension is 35, the audio feature dimension is 74, the dropout rate is 0.1, and the number of heads in multi-head attention is set to 5. The number of stacked layers in the attention layer is set to 5.

C. Baselines

To assess the relative performance of the TGMoE model, we will compare it with the following baseline models.

TFN [1]: Tensor fusion network decomposes unimodal vectors into tensors through the Cartesian product, then fuses the outer product of tensors to learn interactions within and between modes.

LMF [6]: Efficient low-rank multimodal fusion decomposes stacked high-order tensors into many low-rank factors, then efficiently fuses them based on these low-rank factors to improve efficiency.

MFM [20]: Multimodal representation learning factors link a multi-modal discriminative network with a generative network possessing intermediate modality-specific factors to facilitate the reconstruction of the fusion process and optimize the discriminative loss.

MulT [16]: Multimodal Transformer constructs a Transformer between modalities through attention mechanisms, integrating multimodal information and optimizing the fusion process.

MAG-BERT [12]: The multimodal adaptive gate designs an alignment gate for integrating visual and audio information and integrates it into a standard BERT model.

MISA [21]: The modality invariance and specificity of multimodal sentiment analysis project features into two separate independent spaces with specific constraints, taking into account the invariance and specificity of modalities, and then complete fusion on the features of both spaces.

BIMHA [22]: Enhancing bimodal information for arbitrary pairs of modalities using a multi-head attention mechanism, utilizing tensor fusion to capture interactions between modalities, effectively integrating information carried by different modalities, and improving sentiment prediction.

SELF-MM [23]: Self-supervised multi-task learning assigns a single-modal training task with self-generated labels to each modality, aiming to learn the consistency between modalities and the specificity within each modality.

CubeMLP [24]: By mixing relevant modality features on three axes using three independent MLP units and flattening the mixed multimodal features for task prediction.

TETFN[4]: By utilizing visual features extracted by the Vision Transformer, combined with audio features, learning text-oriented cross-modal mappings in pairs, in order to obtain efficient multimodal representations for emotion prediction.

D. Results

1) Comparison experiment with the baseline model: Tables I and II provide the experimental results of various models on the CMU-MOSI and CMU-MOSEI datasets. For Acc-2 and F1 values, there are two sets of evaluation results: on the left, positive and neutral sentiment samples are considered positive examples, and negative samples are considered negative examples to calculate accuracy. On the right, positive sentiment samples are considered positive examples, and negative samples are considered negative examples to calculate accuracy. Similarly, F1 values are calculated accordingly to obtain the corresponding F1 scores.

It can be seen that the proposed TGMoE model achieved the best performance on both datasets. On the CMU-MOSEI dataset, the Acc-2 and F1 scores were improved by 1.11%/0.33% and 1.4%/0.59%, respectively compared to previous methods. On the CMU-MOSI dataset, Acc-7, Acc-2 (left), and F1 scores (left) were improved by 0.1%, 1.32%, and 1.4% compared to previous methods. This indicates that the TGMoE model can adequately integrate different modalities of sentimental information, enhance non-textual modality

sentimental information to contribute more to textual modality sentiment representation in sentiment analysis, and effectively balance the semantic gap between different modalities.

TABLE I. EXPERIMENTAL RESULTS OF THE TGMOE MODEL AND BASELINE MODELS ON THE CMU-MOSI DATASET

Model	MAE	Corr	Acc-7	Acc-2	F1
TFN	0.901	0.698	34.90	-/80.80	-/80.70
LMF	0.917	0.695	33.20	-/82.50	-/82.40
MFM	0.877	0.706	35.40	-/81.70	-/81.60
MulT	0.861	0.711	-	81.50/84.10	80.60/83.90
MAG-BERT	0.727	0.781	43.62	82.37/84.43	82.50/84.61
MISA	0.804	0.764	-	80.79/82.10	80.77/82.03
BIMHA	0.925	0.671	36.44	78.57/80.30	78.57/80.30
SELF-MM	0.712	0.795	45.79	82.54/84.77	82.68/84.91
CubeMLP	0.770	0.767	45.50	-/85.60	-/85.50
TETFN	0.717	0.800	-	84.05/86.10	83.83/86.07
TGMoE	**0.760**	**0.767**	**45.89**	**85.37/85.64**	**85.23/85.71**

TABLE II. EXPERIMENTAL RESULT OF THE TGMOE MODEL AND BASELINE MODELS ON THE CMU-MOSEI DATASET

Model	MAE	Corr	Acc-7	Acc-2	F1
TFN	0.593	0.700	50.20	-/82.50	-/82.10
LMF	0.623	0.677	48.00	-/82.00	-/82.10
MFM	0.568	0.717	51.30	-/84.40	-/84.30
MulT	0.580	0.703	-	-/82.50	/82.30
MAG-BERT	0.543	0.755	52.67	82.51/84.82	82.77/84.71
MISA	0.568	0.724	-	82.59/84.23	82.67/83.97
BIMHA	0.559	0.731	52.11	84.07/83.96	83.35/83.50
SELF-MM	0.529	0.767	53.46	82.68/84.96	82.95/84.93
CubeMLP	0.529	0.760	54.90	-/85.10	-/84.50
TETFN	0.551	0.748	-	84.25/85.18	84.18/85.27
TGMoE	**0.535**	**0.757**	**53.70**	**85.36/85.51**	**85.58/85.86**

2) Ablation studies: In the field of sentiment analysis, compared to audio and visual modalities, the text modality contains more sentimental semantic information. Therefore, the model proposed in this paper is text-centric, where the sentimental information from audio and visual modalities is extensively extracted and integrated into the text modality. Subsequently, a sparsely gated MoE network is utilized to select different expert combinations to analyze and process the sentimental information.

To validate the rationality of the proposed fusion approach, we further explored the impact of different modalities and sparsely gate MoE on sentiment analysis results of the CMU-MOSEI dataset. Experimental results are shown in Tables III and IV, where T, V, and A, respectively represent text, visual, and audio modalities. TGMoE-NoMoE indicates the TGMoE model without the hybrid expert module, where the extracted multimodal features are directly connected to a fully connected layer for sentiment prediction. This leads to the model being unable to learn sufficient nonlinear representations from multimodal features. TGMoE-FFN represents the TGMoE model replacing the hybrid expert module with an FFN layer, where the multimodal features output by the text guided cross-modal feature fusion module are input into the FFN layer to strengthen the multimodal features. This results in the model only being able to learn limited sentimental information, directly impacting the accuracy of sentiment analysis results.

TABLE III. EXPERIMENTAL RESULTS OF ABLATION EXPERIMENTS INVOLVING DIFFERENT MODALITIES IN FUSION ON THE CMU-MOSEI DATASET

Num	Model	MAE	Corr	Acc-7	Acc-2	F1
1	A	0.839	0.012	41.36	71.02/62.85	83.06/77.19
2	V	0.810	0.217	41.92	75.74/70.77	67.85/60.84
3	T	0.589	0.713	50.16	80.92/82.77	80.82/83.13
4	A+V	0.810	0.229	41.39	64.99/63.43	65.41/65.15
5	T+A	0.575	0.720	51.86	79.87/84.31	79.29/84.35
6	T+V	0.584	0.703	50.91	82.08/83.41	82.17/83.89
7	A+V+T	0.535	0.757	53.70	85.36/85.51	85.58/85.86

TABLE IV. EXPERIMENTAL RESULT OF THE EFFECTIVENESS EXPERIMENT OF MOE MODULE ON THE CMU-MOSEI DATASET

Num	Model	MAE	Corr	Acc-7	Acc-2	F1
1	TGMoE-NoMoE	0.587	0.714	50.88	81.28/83.73	82.13/83.65
2	TGMoE-FFN	0.564	0.729	52.09	83.62/84.36	83.65/84.58
3	TGMoE	0.535	0.757	53.70	85.36/85.51	85.58/85.86

E. Discussion

From the perspective of the number of modalities, information from each modality can complement each other, and the addition of more modal information will lead to better sentiment analysis results. For the single modality baseline, the performance of the text modality is superior to the audio and visual modalities, indicating that the text modality contributes more to multimodal sentiment analysis than the audio and visual modalities. In real life, people also prefer to use language to express their direct sentiments.

From the perspective of model modules, when the MoE module is removed, all metrics of the model are lower compared to the complete model. Substituting the MoE module with an FFN (Feedforward Neural Network) layer leads to an improvement in model performance compared to TGMoE-NoMoE, indicating that FFN can strengthen the sentimental information in multimodal features, but with limited performance. The complete TGMoE model outperforms all metrics. Experimental results suggest that MoE plays a crucial role in multimodal fusion, and removing or simply replacing MoE results in the model's inability to learn sufficient nonlinear representations from multimodal features, thereby failing to guarantee that the final fused features contain abstract sentimental information, consequently affecting sentiment prediction performance.

In conclusion, the experiments above validate the efficacy of the text guided mixture of experts model TGMoE. This model adeptly harnesses sentimental data across various modalities, facilitating efficient fusion, mitigating the impact of sentimental information discrepancies across modalities on the final sentiment, and bolstering the efficacy of multimodal sentiment analysis.

V. CONCLUSION

To address the issue that the obtained multimodal fusion representation may be defective in capturing sentimental information due to ignoring the different contributions of various modalities to sentiment analysis, this paper proposes a

text guided mixture-of-experts model TGMoE for multimodal sentiment analysis. The TGMoE model is structured around three key modules. Firstly, features are extracted for each of the three modalities respectively to capture the inherent consistency within each modality. Secondly, a text guided cross-modal fusion mechanism is proposed: cross-modal attention mechanisms are used for text-visual and text-audio modalities, respectively, to capture the interactive information of visual and audio modalities with the text modality, supplementing the text modality with the sentimental information from the visual and audio modalities. Finally, a sparsely gated mixture of expert networks is employed to fortify the nonlinear representational capacity within multimodal features, engender more abstract fusion features, and elevate the precision of sentiment polarity classification. Comparative evaluations against existing multimodal sentiment analysis frameworks demonstrate a pronounced performance boost, underscoring the efficacy of the proposed text guided cross-modal interactive approach and the utility of employing mixture of expert networks for sentiment analysis enhancement.

In real-world scenarios of multimodal sentiment analysis, users may not provide information from all modalities simultaneously. For example, they might only provide text while missing audio or visual data. Therefore, our future research will focus on effectively handling cases of missing modalities, developing more robust sentiment analysis models, and enhancing the practical application value and user experience of these models.

ACKNOWLEDGMENT

This work was supported the National Natural Science Foundation of China (Grant No. 71473034), and the Heilongjiang Provincial Natural Science Foundation of China (Grant No. LH2019G001).

REFERENCES

[1] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[2] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Computational Intelligence*, vol. 36, no. 2, pp. 861–881, 2020.

[3] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, p. 1823.

[4] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[5] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis," *Information Fusion*, vol. 66, pp. 184–197, 2021.

[6] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2247–2256.

[7] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 481–492.

[9] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1009–1021.

[10] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5666–5675.

[11] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.

[12] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020, 2020, p. 2359.

[13] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021.

[14] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing*, 2014, pp. 960–964.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[17] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.

[18] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[19] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[20] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *International Conference on Representation Learning*, 2019.

[21] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[22] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, p. 107676, 2022.

[23] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[24] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 3722–3729.