IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 October 2024, accepted 4 November 2024, date of publication 8 November 2024, date of current version 18 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3493933

RESEARCH ARTICLE

Multimodal Sentiment Analysis of Government Information Comments Based on Contrastive Learning and Cross-Attention Fusion Networks

GUANGYU MU[1,2], CHUANZHI CHEN[1], XIURONG LI[3], JIAXUE LI[1], XIAOQING JU[1], AND JIAXIU DAI[1]

[1]School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China
[2]Key Laboratory of Financial Technology of Jilin Province, Changchun 130117, China
[3]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Corresponding author: Xiurong Li (xiurong_li@bjut.edu.cn)

This work was supported in part by the National Social Science Fund of China under Grant 19BJY246, in part by the Natural Science Fund Project of the Science and Technology Department of Jilin Province under Grant 20240101361JC, and in part by the Think Tank Fund Project of the Jilin Science and Technology Association.

ABSTRACT Accurate identification of sentiments in government-related comments is crucial for policy-makers to deeply understand public opinion, adjust policies promptly, and enhance overall satisfaction. Thus, we create a model for emotion recognition in multimodal sentiment analysis of government information comments based on contrastive learning and cross-attention fusion networks. Firstly, we collect text-image comments from Today's Headlines App's Politics and Law section and extract textual and visual features. We fine-tune the model with LoRA and optimize the feature representation by making low-rank adjustments to the fused features. Secondly, we utilize contrastive learning with reverse prediction to analyze intra-class and inter-class cross-modal dynamics. Then, we propose a novel fusion network that utilizes cross-attention to learn the complementary relationship between different modalities. Finally, the features are combined using the fully connected layer. The experiment illustrates that the model achieves a 96.80% accuracy in recognizing emotion polarity. Compared with the multimodal model CLIP, the accuracy of the proposed method is improved by 10.21%. The model could assist the government in emotional evolution analysis, detection of public opinion, and online public opinion guidance.

INDEX TERMS Government information comments, multimodal sentiment analysis, contrastive learning, cross-attention, fusion networks.

I. INTRODUCTION

With the rise of mobile internet and smartphones, social media platforms like Twitter, microblogs, and news apps have become primary sources for accessing current affairs [1]. Posts related to government information often receive high engagement on social media platforms. Government information refers to official data released by governments, enterprises, institutions, or authoritative media organizations. The comments section on online platforms frequently attracts much attention, leading to many comments from internet users [2]. These commenters come from diverse backgrounds and share a common interest. They express and discuss their opinions in the comment section [3]. As a result, comments on social media are more likely to influence online public opinion regarding perspectives and emotions than the underlying topic [4]. Online public sentiment triggered by government information spreads rapidly and has a broader impact than general public sentiment. The government must promptly and correctly address the situation to prevent negative emotions among internet users. Noncompliance can have harmful consequences for government legitimacy and society [5]. Therefore, performing a sentiment analysis of government information is crucial to

The associate editor coordinating the review of this manuscript and approving it for publication was Loris Belcastro.

ensure that the public perceives the government's message positively.

The field of sentiment analysis has two stages: unimodal and multimodal. Scholars initially focus on the emotional analysis of textual language [6]. For instance, the SCA-HDNN proposed by Khan et al. [7] builds CNN and hybrid DNN networks and performs emotional analysis of text reviews in different fields such as film, banking, e-commerce, and social media. However, analyzing and expressing views and emotions online has become more challenging due to the increasing use of multimodal content, including text and images [8]. Scholars have explored alternative modalities of data to obtain more comprehensive emotional information [9]. To carry out sentiment classification by combining data from multiple sources, Shaik et al. [10] proposed multimodal fusion methods that utilize machine and deep learning technologies. These methods enable models to improve the accuracy of sentiment analysis by learning associations between additional modal data. For example, Shi et al. [11] used CNN and attention mechanisms to construct the MSCNN-CPL-CAFF model and analyze the public's attitude to COVID-19 on a short video platform. Multimodal sentiment analysis has shown a significant improvement in prediction accuracy compared to unimodal sentiment analysis.

However, many previous studies have required assistance to achieve accurate alignment between different modalities due to the inherent heterogeneity of multimodal data [12]. This heterogeneity manifests in variations of modalities, including significant differences in data distribution, dimensions, and modes of expression. Past research has focused chiefly on interactive learning within the same category, also known as intra-modal learning in a sample-in setting. When we deal with multimodal sentiment data, it is essential to consider the relationships between samples from different categories, including their dynamic interactions. Although some methods have been developed to enhance the alignment of modalities [13], they often need help to fully capture the relationships within different categories, which limits the model's performance [14]. In contrast, contrastive learning has proven to be a more practical approach for addressing the issue of modal alignment. Contrastive learning enables the model to comprehensively grasp the dynamic relationships within and between classes by learning samples' similarities and differences. Contrastive learning emphasizes samples' global emotional relationships, including intra-modal relationships within the same category and inter-modal relationships of different categories. Therefore, contrastive learning is a better approach to modal alignment in multimodal sentiment analysis.

Although natural language plays a crucial role in determining sentiment, other modes of information can also provide additional emotional semantic cues [15]. Scholars have explored various methods of integrating multimodal information to strengthen task understanding. For instance, the feature concatenation method directly combines features from other modalities into longer feature vectors [16]. When combining different modalities, it is vital to manage them effectively to prevent information loss or redundancy. Alternatively, feature-weighted fusion involves assigning weights to each modality and combining the weighted modal features. If the label information is missing or incorrect, it can seriously affect the results of weight distribution. Attention networks have been incorporated into multimodal sentiment analysis to extract emotion-related information from multiple modalities [17]. For example, Feng et al. [18] used attention mechanisms on the critical time steps of emotional change in an audio module. Li et al. [19] designed a cross-modular focus mechanism to capture long-term dependencies of elements between different modular inconsistencies. Extensive research has been conducted on the attention mechanism for combining multiple modes of information. However, most studies overlook the dynamic changes in modalities due to their differences. In general, information from different modalities has complementary characteristics. Cross-attention networks are a common attention mechanism in multimodal tasks or multitask learning. They establish connections between different input sequences. Using cross-attention, the model can identify relevant areas of an image and corresponding text. Cross-attention allows the model to comprehend the connections between various input sequences better. As a result, cross-attention is a valuable technique for combining government information in text and image formats.

Therefore, this paper raises the following questions.

1) How can we conduct a multimodal emotional analysis of comments on government information?

2) How can we entirely use inter and inter-class cross-modal dynamics to respond to the specificity of the government information?

3) How can we solve the problem of the fusion of multimodal emotional features?

In response to the above questions, we create a model to analyze text-image comments on government information and propose a feature fusion strategy that utilizes cross-attention to combine text and image features. Our training process utilizes contrastive learning to maximize the use of the limited dataset and improve the model's performance. The final results of sentiment analysis can be used to analyze sentiment evolution, detect public opinion, and guide online public opinion.

The paper makes the following contributions.

1) The paper creates a multimodal sentiment analysis architecture for analyzing government social media content sentiment.

2) The paper designs cross-modal contrastive learning for training samples' intra-class and inter-class cross-modal dynamics.

3) The paper proposes a feature fusion network that uses cross-attention and fuses sentiment analysis results of different modalities through decision fusion.

The structure of this paper is as follows. Section II provides an overview of the related research. Section III presents a novel multimodal sentiment analysis model for government information comments. Section IV conducts experiments and results. Section V is the conclusion and prospect of this paper.

II. RELATED WORK

A. UNIMODAL SENTIMENT ANALYSIS

Sentiment analysis techniques can be classified into three categories, including sentiment dictionaries [20], machine [21], and deep learning [22] methods.

During the early stages, sentiment analysis primarily relied on sentiment dictionaries [23] to identify the emotional tone by comparing vocabulary in dictionaries and corpora. For instance, Mu et al. [24] studied comments about various stocks using a sentiment dictionary. They predicted stock prices by tracking changes in investor sentiment. Despite the method can quickly analyze a text's sentiment, researchers must continuously add polysemous words due to the constant emergence of new words and ambiguity. Therefore, scholars have begun to use machine learning algorithms for sentiment analysis.

Support Vector Machine (SVM) is widely used in emotional analysis. Han et al. [25] proposed a method to address the issue of hidden emotional features in text analytics. Specific machine learning algorithms have remarkably succeeded in accurately categorizing human emotions. The study by Piryani et al. [26] deeply explored the effects of machine learning in sentiment analysis tasks. With the introduction of ensemble learning, methods such as decision-making trees begin to play a role in sentiment analysis. Kazmaier and van Vuuren [27] summarized the accomplishments of ensemble learning in their review study. These achievements provide a solid theoretical foundation and practical experience in sentiment analysis. Although machine learning-based methods have improved the efficiency of sentiment analysis, they typically rely on large amounts of labeled data. Moreover, machine learning-based methods are noise-sensitive and struggle to handle long-range dependencies.

Deep learning-based methods transform text into vector representations and build neural network models to capture context semantic information for predicting sentiment inclination. Fan et al. [28] demonstrated that the SDCNN model, based on a convolutional neural network with sparse dropout, outperformed the CNN in classification performance. These deep learning techniques are helpful for future tasks that deal with complex and diverse emotional data. Deep learning-based methods have achieved good results. However, they treat text as a sequence of words to obtain embeddings, which cannot capture the intra-class relationships of different samples.

Recently, scholars have researched government information that mainly focuses on two aspects. On the one hand, the focus is on studying the public's opinion regarding government information. Zhang et al. [29] analyzed the state of the government's information public opinion governance and explored how to manage and guide public opinion while presenting strategies and recommendations for future regulation. Chen et al. [30] investigated the relationship between themes and user behaviors in government short videos on the TikTok platform. At the same time, Lei et al. [31] explored Weibo's public opinion on a specific topic. On the other hand, research is conducted on public and government communication mechanisms. Rahmanti et al. [32] suggested that effective communication of public opinions on social media can foster people's trust in the government. Lerouge et al. [33] developed new methods for monitoring public emotional responses on social media to support government communication during crises. The above studies either focus on unimodal information or emphasize communication mechanisms. Our research addresses multimodal comments on government information.

B. MULTIMODAL SENTIMENT ANALYSIS

Accurately judging emotions from real-world experiences requires multimodal learning involving text and images. Multimodal emotion analysis is based on unimodal emotional analysis involving simple connections or complex deep neural networks (DNNs) [34], [35]. Some scholars have also sought to enhance multimodal characterization by introducing other aspects of emotional information, which helps obtain richer modular characteristics and supports more accurate emotional analysis. Zou et al. [36] extracted more profound emotional clues from the conversation by setting up an emotional hint extractor. Zhang et al. [37] used unmarked audio-visual data to help learn speech characteristics in speech-emotional analysis based on transfer and semi-surveillance learning. Jin et al. [38] combined the Oxford Dictionary's external word interpretation knowledge with the enhanced aspects-based emotion analysis methodology.

It is challenging to extract and express textual features from online public opinion stirred up by government information. This semantic complexity poses difficulties for emotion recognition. Most studies in this area have conducted sentiment analysis on a single modality or the main content of posts. However, the systematic research and summarization of multimodal sentiment analysis on government information commentary is insufficient.

However, these methods are simple and rough when modeling the cross-modal semantic space. They ignore relationships between different classes and do not availably combine different types of information. As a result, the methods cannot effectively integrate complementary and multi-level multimodal information to extracting high-quality multimodal feature representations. Therefore, we propose

using contrastive learning. Contrastive learning is efficient at capturing the dynamic relationships within and between classes. It enhances the model's understanding of multimodal features and improves task accuracy. We employ the principles of contrastive learning at the feature extraction stage to obtain more consistent modal features.

C. MULTIMODAL FUSION STRATEGIES

In contrast to unimodal sentiment analysis, the most crucial step in multimodal sentiment analysis after feature extraction is the fusion of features from different modalities. The selection of appropriate feature fusion techniques affects the accuracy of multimodal sentiment analysis. Modal fusion can be classified into three types based on modal fusion and modeling sequence: feature-level fusion, decision-level fusion, and hybrid fusion [39], [40].

Feature-level fusion combines features from different modalities to create a unified feature vector for sentiment analysis tasks. Before the fusion process begins, the features extracted from other modalities are converted to the same format and then spliced. This method is extensively used in various sentiment analysis tasks. Zadeh et al. [39] first introduced the assignment of trimodal sentiment analysis and utilized automatic extraction of multimodal features for sentiment analysis. They automatically identified emotional cues in the spoken text to generate discourse features and extracted visual and audio features from video sequences. After concatenating the extracted features from each modality, a trimodal HMM classifier was utilized to learn the input signals' hidden structure. Pérez-Rosas et al. [41] proposed a discourse-level sentiment analysis and constructed the first MOUD dataset. The method built vocabulary tables, used simple weighted graph features as textual sentiment features, and utilized OpenEAR for audio feature extraction and CERT for facial feature extraction. Then, they merged the features and employed a SVM classifier to ascertain the polarity of the sentiment. However, an obvious drawback of feature-level fusion is the potential loss of modality-specific nuances when combining features into a unified representation.

Decision-level fusion, also known as late fusion, conducts sentiment analysis for each modality and then incorporates the results of unimodal sentiment analysis into different mechanisms for the final decision. This approach enables each modality to utilize its most fitting classifier to learn its features. However, each modality has its classifier, and the interaction between modalities is often challenging to model effectively, making the learning process of classifiers complex and time-consuming. Nojavanasghari et al. [42] developed a unimodal classifier capable of handling video, audio, and text models. Then, they computed the average confidence of each unimodal classifier to predict the final result. Wang et al. [43] created a SAL-CNN model to improve its generalizability and its ability to predict emotions. The SAL method included two stages: selection and addition. During the selection stage, the model analyzed the latent representations produced by the neural network to identify confounding factors. The model eliminated these confounding elements in the addition stage by adding Gaussian noise to these representations. Nevertheless, decision-level fusion needs more valid interaction modeling between modalities, resulting in a less robust cross-modal understanding.

A robust multimodal data fusion scheme can extract and integrate critical information from various sources while preserving the interdependence of those sources. One approach is hybrid fusion, combining feature fusion and decision fusion to compensate for their shortcomings. For example, Hussain et al. [44] proposed a hybrid fusion method based on weighted majority voting to complete sentiment classification. Zhang and Jiang [45] established the FCNN model to fuse text emojis and picture feature vectors extracted at the feature layer and identify the presence of sarcasm in multimodal tourism comments. While compensating for individual fusion shortages, hybrid fusion's complexity can increase computational overhead and cause difficulty balancing features and decision integration.

With the rise of attention networks [46], fusion methods based on attention have become the predominant approach in multimodal sentiment analysis. Tsai et al. [47] proposed a multimodal transformer called MulT. It uses a cross-channel attention interaction module to highlight interactions between multiple channels simultaneously. Hazarika et al. [48] developed the MISA model framework to integrate information about interactions within and between different modalities by mapping each modality to its own space and a cross-modal shared space. Zhang et al. [37] proposed a bidirectional masked attention mechanism for text and speech bimodal. This mechanism dynamically incorporates information from another modality using masked attention, assigns attention weights to the current modality, and minimizes the disparities between modalities.

Our proposed fusion strategy based on a cross-attention network shows better information complementarity in feature-level and decision-level fusion. It improves the overall performance of emotion recognition effectually. Compared with the existing methods, our method significantly enhances the sensitivity and accuracy of the model by introducing the learning of inter-class and intra-class relations.

III. METHODOLOGY

In this section, we introduce the methods used to design the model, including feature extraction and representation, model fine-tuning with LoRA, the concept of contrastive learning, and feature fusion using cross-attention.

A. FEATURE EXTRACTION AND REPRESENTATION

We use BERT and CLIP to extract text and image features based on the Transformer and obtain an optimal representation of multimodal features.

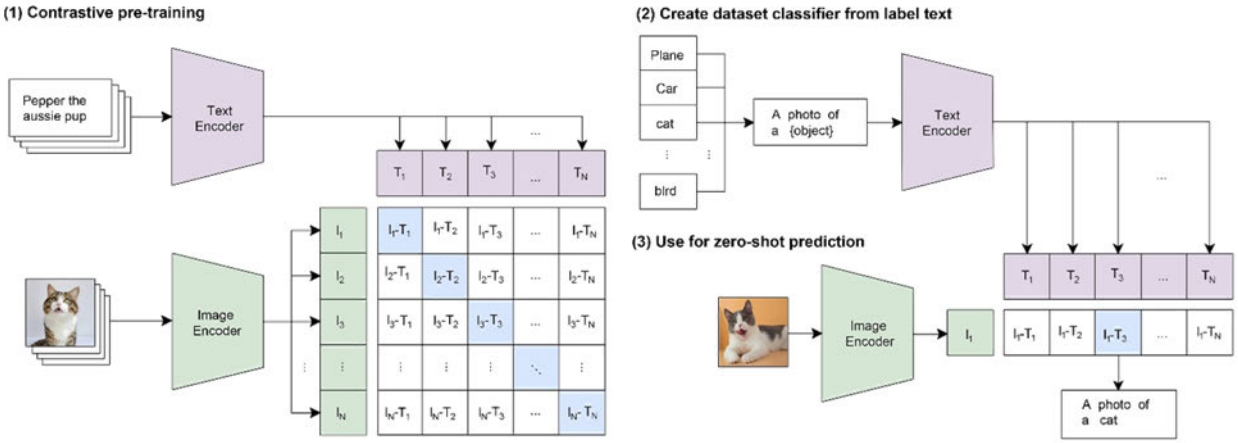

FIGURE 1. The structure of the CLIP model.

1) TEXT FEATURE EXTRACTION & REPRESENTATION

NLP tasks face significant challenges in representing text features as text complexity increases. To tackle this issue, the Transformer model utilizes self-attention to calculate input and output representations. The model can transform one sequence into another using an encoder and decoder. The two primary functions of the Transformer-based model BERT are Next Sentence Prediction (NSP) and Masked Language Model (MLM). BERT is a valuable tool for extracting features that can overcome the limitations of traditional neural network models in understanding contextual semantic relationships. Considering the intricate semantic nature of public opinion that arises from government information is beneficial. Extracting and representing text features can be challenging, but BERT can handle it effectively. We utilize the BERT basic model with 12 self-attention heads to extract text features and get a vector dimension of 768. The input sequence T is constant, and the BERT model generates the contextual representation of each word in the sequence.

2) IMAGE FEATURE EXTRACTION & REPRESENTATION

Traditional image classification models encounter numerous limitations when handling large-scale image classification tasks. These models need more annotated data to generalize to new image categories. However, the CLIP model is a pre-trained model that helps understand semantic links between images and texts. Simultaneously, CLIP learns generalized characteristic expressions through self-supervised contrastive using large amounts of unmarked image and text data.

An image and text encoder processes images and texts separately and maps them into a joint semantic space as shown in Figure 1. CLIP maximizes the similarity between matched pairings and minimizes the similarity between non-matching pairs to optimize the model. The contrastive learning method allows CLIP to capture rich visual and linguistic features and their associations. After completing pre-training, CLIP performs zero-shot prediction and forecasts the new task directly. The process involves analyzing the image coding and comparing the text description codes in the semantic space. The model then selects the text description that is most similar to the predicted result. Thus, the new categories of images are classified and understood without additional training data. Our image processing model utilizes the Vision Transformer (ViT), where the image is set as I and the output of the CLIP model is the semantic representation of the image H_i.

B. MODEL FINE-TUNING BASED ON LoRA

This paper uses LoRA (Low-Rank Adaptation) to fine-tune our multimodal sentiment analysis model. LoRA is a technique that addresses the high computational costs large pre-trained language models. The weights of a pre-trained model are frozen. Each Transformer block now includes trainable layers composed of rank-decomposed matrices. The GPU memory is also reduced by decreasing the number of trainable parameters.

Equations 1 and 2 indicate that the LoRA attention mechanism effectually captures the connections between different modalities and increases the model's comprehension of the whole dataset. We improve the unimodal representations by linking text and image and adjusting attention matrices A_t and A_i to enhance local relevance. $H_m^{\text{finetuned}}$ represents the fine-tuned text or image feature. A_m is the attention matrix for text or image modules that adjusts the original text or picture feature H_m. The adjustment enhances the model's data processing capabilities, particularly the correlation between different modalities. We use BERT and CLIP to extract text and image features based on the Transformer and obtain an optimal representation of multimodal features.

$$H_t^{\text{finetuned}} = A_t \cdot H_t \tag{1}$$

$$H_i^{\text{finetuned}} = A_i \cdot H_i \tag{2}$$

C. CONTRASTIVE LEARNING WITH REVERSE PREDICTION

Our research involves an innovative approach to contrastive learning, which enhances the interaction between different data types through Contrastive Predictive Coding (CPC). It is a self-supervised learning technique commonly used in audio and image analysis. By employing contrastive learning, CPC introduces self-supervised tasks that enable the model to learn the data structure without relying on explicit labels.

In this model, contrastive learning plays a crucial role in improving the accuracy of sentiment analysis by capturing the dynamic changes within and between classes. Intra-class dynamics refers to the relationships and variations between samples within the same class, such as positive reviews. In contrast, inter-class dynamics refers to the differences between different classes, such as positive and negative reviews. Intra-class relations enable the model to better understand and strengthen the connection between similar samples by comparing features in samples of the same class. The inter-class relationship promotes the model to distinguish the differences between different classes by introducing negative sample comparison to improve classification accuracy. This dual focus allows the model to learn subtle sentiment representations, enhancing its ability to classify reviews accurately.

Our model uses CPC to measure the mutual information between contextual and future elements. This approach effectively reduces high-dimensional data into a condensed hidden space for more straightforward conditional prediction. The method enables the fused result to capture more modal-invariant information by predicting the representation across modalities in reverse order. Additionally, the model can adjust its predictions for each modality to determine the necessary information to improve its flexibility.

To implement this method, we define the fused feature as $\hat{X}_o$ and the features of two modalities as h_t and h_v. We quantify their correlation by applying a scoring function to the normalized prediction and actual value vectors. The scoring function is presented in Equations 3, 4, and 5.

$$\overline{G_{\emptyset}\left(\hat{X}_o\right)} = \frac{G_z\left(\hat{X}_o\right)}{\left\|G_z\left(\hat{X}_o\right)\right\|_2} \tag{3}$$

$$\overline{h_m} = \frac{h_m}{\|h_m\|_2} \tag{4}$$

$$\mathrm{s}\left(h_m, \hat{X}_o\right) = exp\left(\overline{h_m}\left(\overline{G_z\left(\hat{X}_o\right)}\right)^T\right) \tag{5}$$

In this function, $G_{\emptyset}$ represents a neural network composed of parameters $\emptyset$, which calculates the distances of the fused feature, and h_m represents the distance of the molecular characteristics $m \in \{t, v\}$. The fusion feature $\hat{X}_o$ is a combined representation of information extracted from different modules.

All other examples of the modality in the same batch are treated as negative samples by the scoring function, which is part of noise-contrastive estimation. Simultaneously, we introduce contrastive learning loss through Equations 6 and 7 to reduce the distance between modalities for consistency and increase the distance for specificity. The model's prediction determines the amount of information obtained from each modality by adjusting them accordingly.

$$\mathcal{L}_N^{\hat{X}_o, h_m} = -\mathrm{E}\left[\log \frac{\mathrm{s}\left(h_m^i, \hat{X}_o\right)}{\sum \mathrm{s}\left(h_m^j, \hat{X}_o\right)}\right] \tag{6}$$

$$\mathcal{L}_{CPC} = \mathcal{L}_N^{\hat{X}_o, h_t} + \mathcal{L}_N^{\hat{X}_o, h_v} \tag{7}$$

Overall, we optimize the loss function for the entire task L as shown in Equation 8, where $\mathcal{L}_{pred}$ represents the model's prediction error, and α represents the hyperparameter used to adjust the loss function.

$$L = \mathcal{L}_{pred} + \alpha \mathcal{L}_{CPC} \tag{8}$$

Using this method, our model improves its performance in multimodal sentiment analysis tasks and offers a new perspective for comprehending and managing complex interactions between different modalities.

D. MULTIMODAL FEATURE FUSION

This paper introduces an innovative cross-attention network method in multimodal feature fusion. On the one hand, we consider the complementary and associative relationship between text and image. We extract features from these two modalities and perform feature-level fusion using a cross-attention network. On the other hand, decision-level fusion optimizes sentiment analysis by combining text and image recognition in fully connected layers.

1) PROPOSED FEATURE FUSION BASED ON CROSS-ATTENTION

The transformer encoder is connected by a multi-head self-attention (MSA), layer normalization (LN), and a multi-layer perceptron block applied with residual connections. The definitions are as follows.

$$y^l = MSA\left(LN\left(h_m^l\right)\right) + h_m^l, m \epsilon \{t, v\}, X_m \epsilon R^{l_m \times d_m} \tag{9}$$

$$h^{l+1} = MLP\left(LN\left(y^l\right)\right) + y^l \tag{10}$$

$$MSA\left(h_m\right) = Attention\left(W^Q h_m, W^K h_m, W^V h_m\right) \tag{11}$$

In Equations 9, the input of a multimodal sequence $X_m \epsilon R^{l_m \times d_m}$ is encoded as h_m, where l_m is the length of the modal input, and d_m is the dimension of the modal input, with m representing either text or image modality. Equation 10 illustrates the result of encoding one layer of the transformer. y^l is the intermediate result after multi-head attention. As shown in Equation 11, the MSA operation is used to compute the dot-product attention. $\mathrm{W^Q}$, $\mathrm{W^K}$, and $\mathrm{W^V}$ are weight matrices used to linearly project the modal input $\mathrm{h_m}$ to generate queries, keys, and values. These matrices are linear transformations of the same tensor $\mathrm{h_m}$. The purpose

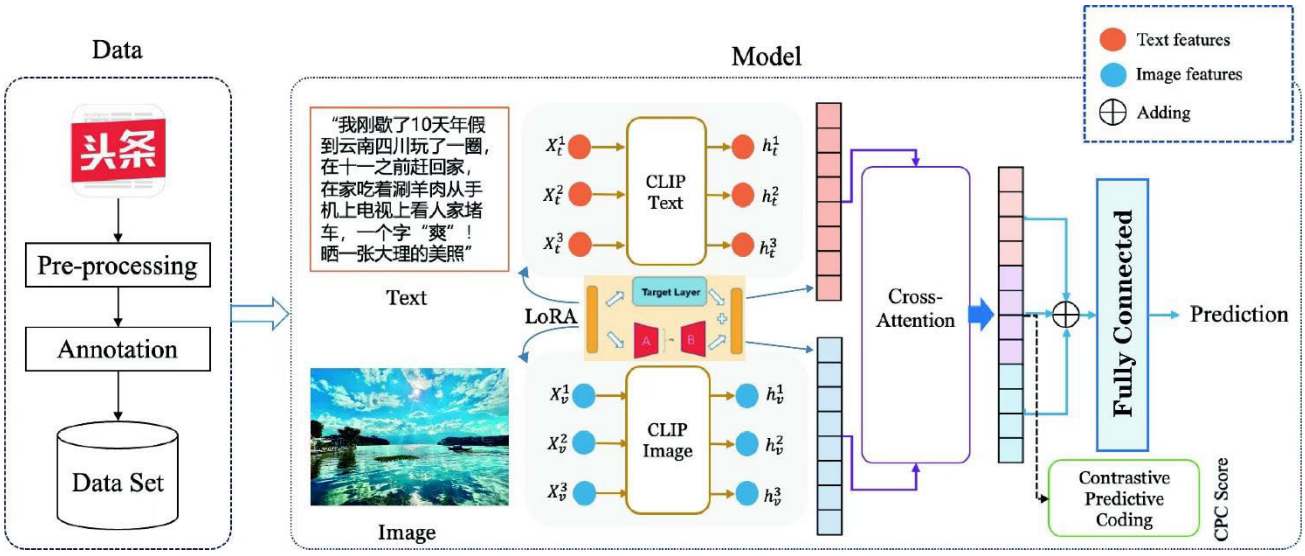

FIGURE 2. Multimodal sentiment analysis model.

of these operations is to characterize the high-level attributes of a text module so that attention is focused more on the characteristics that significantly impact the result. In short, the role of these weights matrices is to linearly transform the original input to generate queries, keys, and values for attention calculations.

In particular, we introduce a novel cross-transformer layer to enhance the interaction between modalities, as shown in Equations 12 and 13. This layer learns directed pairwise attention between source and target modalities to modify the expression of the target modality. For example, text modality information is used to strengthen the expression in the image modality and vice versa. The cross-attention operation improves the information complementarity between modalities for better overall emotion perception and interpretation.

$$h_t^{l+1} = Cross - Transformer(h_t^l, h_v^l; \theta_t) \tag{12}$$

$$h_v^{l+1} = Cross - Transformer(h_v^l, h_t^l; \theta_v) \tag{13}$$

Here, h_t^{l+1} represents the text modality fused with video information, and h_v^{l+1} represents the enhanced image fusion modality. θ denote the model parameters for the text and image modalities, which adjust and optimize the feature representation in the cross-transformation process to enhance the information complementarity between the two modalities. The Cross-Transformer follows the original Transformer operations, with a difference as shown in Equation 14.

$$y^l = MCA\left(LN\left(h_{v,t}^l\right), LN\left(h_{t,v}^l\right)\right) + h_{v,t}^l \tag{14}$$

The fusion principle is illustrated in Figure 2, where modality β represents features extracted by CLIP and modality α refers to features extracted by BERT. X_m indicates the modality feature representation. T_m and d_m represent the sequence length of that modality data and the dimension of each modal input, with m being either text or image modality. W^Q, W^K, and W^V are weight matrices. The softmax is an activation function that normalizes attention weights, ensuring the sum of all attention scores equals 1.

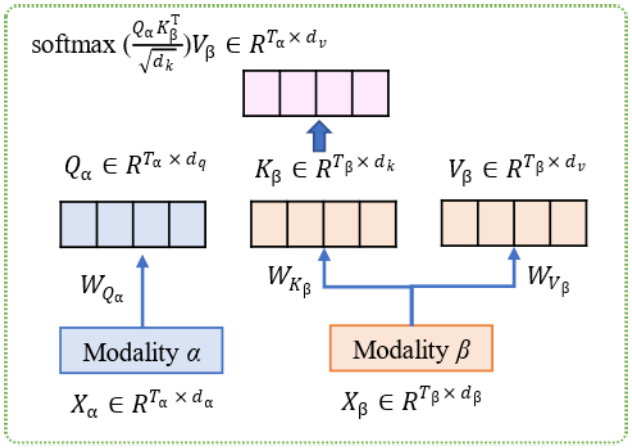

FIGURE 3. Principle of cross-attention fusion.

2) DECISION FUSION BASED ON CROSS-ENTROPY LOSS

In the multimodal sentiment recognition, we employ a decision fusion strategy based on cross-entropy loss. This strategy involves merging features from text and images and then combining these multimodal features to make the final sentiment prediction. We utilize a fully connected network of two layers to accomplish our goal. The method integrates information from multiple modalities at the decision-making level to achieve more precise sentiment predictions. The procedure is implemented in Equations 15 and 16.

$$\hat{y}_i = W_2\left(ReLU\left(W_1\left(\left[h_f, h_a, h_t\right]\right) + b_1\right)\right) + b_2 \in R^{d_{out}} \tag{15}$$

$$\mathcal{L}_{pred} = CrossEntropy\left(y_i, \hat{y}_i\right) \tag{16}$$

Here, ReLU is an activation function that introduces non-linearities, enabling the model to learn complex feature representations better. y represents the accurate sentiment label, and the cross-entropy loss $\mathcal{L}_{pred}$ measures the difference between the two losses. W_1 and W_2 represent the weights, b_1 and b_2 represent the bias terms, and d_{out} indicates the dimension of the output categories.

E. MODEL CONSTRUCTION

This paper proposes an innovative multimodal sentiment analysis model for governmental information comments, as illustrated in Figure 3. First, the model processes text and image data using dataset annotations. Then, the BERT model extracts text features, the CLIP model extracts image features, and a feature vector is obtained. The cross-attention network integrates text and image information by learning the complementary relationships between different modalities. LoRA technology is applied to adjust the low-ranking fusion features to optimize feature representation and enhance the model's sentiment analysis. After integrating and fine-tuning LoRA features, the text and images are merged to classify sentiment using a fully connected layer.

IV. EXPERIMENT

A. EXPERIMENTAL SETUP AND EVALUATION METRIC

The hyperparameter settings for the experiment are shown in Table 1. The entire model is optimized using the Adam optimizer, updated at a learning rate 1e-5, with a batch size of 4, and trained and tested on a machine equipped with RTX 3090. The experimental dataset uses 80% as the training set and the other 20% as the testing set. Finally, accuracy and F1-score are used as evaluation metrics. True Positive (TP) and True Negative (TN) represent correctly predicted positive and negative cases. False Negative (FN) and False Positive (FP) represent positive and negative cases incorrectly predicted as negative and positive, respectively. The accuracy and F1-score calculation methods are presented in Equations 17 and 18.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{17}$$

$$F1 - score = \frac{2 * TP}{2 * TP + FN + FP} \tag{18}$$

TABLE 1. Experiment hyperparameter settings.

Module Name	Hyperparameter Value
Encoder Layers	4
Encoder Dim	768
Attention Heads	12
α	0.1
Multi-head Attention Dropout	0.4

TABLE 2. Sentiment data annotation results.

Sentiment Polarity	Positive	Neutral	Negative
Number	1677	2078	6676

B. DATA COLLECTION AND PREPROCESSING

Accurately capturing public opinion is crucial for managing government departments, guiding psychological well-being, and controlling public opinion. We select Today's Headlines App as a trustworthy source of information because it includes official accounts and does not categorize unverified user messages. We collect original comments containing both images and text from official accounts such as ''Xinhua Net'' and ''Global Times'' in the ''Politics and Law'' section starting from January 15, 2022, to October 5, 2023. During the data preprocessing, we adopt automated tools to remove apparent noise data. Specifically, duplicate or meaningless text is identified and removed by the Text Cleaner library. For image data, the OpenCV library is used for image preprocessing to delete blurred or low-quality images. We gather 4,431 pairs of text and image data by eliminating single-mode texts, images, and repeated comments. Compared with other data annotation methods, manual annotation is time-consuming and laborious, but it is an accurate and effective method. In this paper, two MLIS master's students annotated all reviews with their sentiment polarity according to the positive, neutral, and negative classification. In the case of inconsistent labels, the two students deliberated among themselves and two professors determine the final emotion label. The results are presented in Table 2.

C. COMPARATIVE EXPERIMENTS

We design four sets of comparative experiments (A, B, C, D) to perform the sentiment classification task to demonstrate the effectiveness of the proposed model in this study. ''T'' and ''V'' represent the textual and image modalities, respectively. The experimental results are shown in Table 3.

1) FEATURE EXTRACTION COMPARATIVE EXPERIMENT

In this experiment, we compare the impact of single-modal and multimodal data on sentiment analysis results. The experiment is labeled A1-A6.

A1: Only use the imagery data available in the dataset.

A2-A5: Select high-performance and interpretable classifiers, including LSTM, GRU, CNN, and BERT, and only use the textual data in the dataset.

A6: Use textual and imagery data available in the dataset.

Firstly, the outcome of A1 is significantly lower than that of other experiments in the same group. The results show that the imaging modality only supports the final sentiment judgment to enhance the information obtained from the text modality.

TABLE 3. Result of comparative experiment.

Group	Module	Model	Accuracy	F1-score
A1	V	ViT	0.3534	0.3428
A2	T	GRU	0.8407	0.8343
A3	T	LSTM	0.8439	0.8366
A4	T	CNN	0.8523	0.8485
A5	T	BERT	0.8673	0.8425
A6	T+V	CLIP	0.8783	0.8675
B1	T+V	Multiplicative β=0.3	0.8113	0.7931
B2	T+V	Multiplicative β=0.5	0.8441	0.8312
B3	T+V	Averaging	0.8648	0.8573
B4	T+V	Concat FC	0.8719	0.8678
B5	T+V	Cross-Attention	0.8914	0.9076
C1	T+V	Cross-Attention + CPC	0.9477	0.9298
C2	T+V	Cross-Attention + CPC + LoRA (Ours)	0.9680	0.9423
D1	T+V	CHFusion [49]	0.7650	-
D2	T+V	bc-LSTM [50]	0.8030	-
D3	T+V	MMMU-BA [51]	0.8231	-

Secondly, A2-A5's results are significantly higher than A1, which indicates a higher predictive accuracy in unimodal emotional analysis than in images. For A2, the GRU model can effectively capture semantic features in text messages and produce accurate emotional classification results. For A3, the LSTM model's efficacy in processing textual data, particularly in emotional analysis tasks that call for taking long-term dependency ties into account. For A4, the CNN model is also effective at handling brief text and extracting local characters in emotional analysis tasks. For A5, BERT achieves the best unimodal classification results due to its complex structure and substantial pre-training data.

Thirdly, the accuracy and F1-score of A6 are slightly higher than that of A2-A5, indicating the importance of considering multimodal data in emotional analysis. This set of experiments reveals the advantages of using a multimodal approach in sentiment analysis. Furthermore, the multimodal approach's effectiveness in capturing more complex emotional information is also certified.

2) FUSION METHODS COMPARATIVE EXPERIMENT

In this set of experiments, we compare the impact of different fusion methods on the sentiment analysis results. The experimental groups are divided into B1-B5.

B1: Set the weight of the text prediction results to 0.3 and the weight of the image prediction results to 0.7.

B2: Set the weight of the text prediction results to 0.5 and the weight of the image prediction results to 0.5.

B3: Add the probabilities predicted by the two modalities together and compute the arithmetic average.

B4: In the early fusion stage, we concatenate the two modalities and then output sentiment classification results through a fully connected layer.

B5: Design feature fusion based on Cross-Attention.

Firstly, the experimental results of B1 and B2 are the lowest among all the experiments conducted. The depth feature interaction is not considered since the B1 and B2 experiments are weighted based on simple prediction results. The complementarity and correlation between the modes cannot be captured effectively when the multimodal features are fused. As a result, partial loss of fusion information affects the model's overall performance.

Secondly, the result of B3 is slightly higher than that of B1 and B2, which indicates that the arithmetic average method can better reflect the fundamental correlation between modalities. However, the method still needs to strengthen modalities' in-depth interaction learning.

Thirdly, the results of B4 are higher than those of the above three experimental groups. The results demonstrate that the modal features can be combined more effectively by early fusion in the feature layer and sentiment classification through the fully connected layer. Using the powerful learning ability of the deep learning model shows significant practicability in multimodal sentiment analysis.

Finally, the accuracy and F1-score of B5 are the highest in this group of experiments. The result further proves that the feature fusion strategy based on cross-attention effectively deals with complex sentiment analysis tasks. The innovative cross-attention mechanism proposed in this paper can understand and explain emotions more comprehensively

by enhancing the complementarity and interaction between modes and the ability of modal integration.

In addition, the results of Group B are generally higher than Group A, which emphasizes the advantages of feature fusion methods in multimodal sentiment analysis. Compared with using only a single modality, fusing text and visual modalities can provide more prosperous and complex sentiment information and improve the accuracy and depth of sentiment analysis.

3) CONTRASTIVE PREDICTIVE CODING (CPC) AND LoRA COMPARATIVE EXPERIMENT

In this set of comparative experiments, our research focuses on assessing the effectiveness of these two methods. We evaluate the impact of contrastive predictive coding and the LoRA method on sentiment analysis results.

Firstly, C1 is higher than B5 in terms of accuracy and F1-score. The comparative results show a significant improvement in the model's accuracy due to incorporating CPC. As a self-supervised learning technique, CPC measures the modalities' mutual information to enhance the ability of the model to capture the relationship between text and image features. The model's enhanced inter-modal dynamic representation enables it to handle complex sentiment analysis tasks more availably, which improves accuracy and F1-score.

Compared to C1, C2 demonstrates an even more significant improvement in accuracy and F1-score, highlighting the importance of LoRA. When tuning large language models, LoRA reduces the number of parameters that need to be trained through low-rank decomposition, enabling the model to adapt to specific tasks and fine-tune models more efficiently. LoRA helps a model better capture and comprehend the correlation and reciprocal information between text and images, which is crucial for sentiment analysis, by varying the weight of the attention mechanism.

4) EXPERIMENTAL RESULTS COMPARISON

We select three baseline models in Group D to evaluate the relative superiority of our model. Our proposed model has shown a significant improvement in accuracy compared to Group D models. The use of CPC reinforces the gap between different samples. Moreover, our fusion method effectively retains the features between different modalities. The LoRA-based fine-tuning technique also reduces the computational power and time costs associated with training the model.

The comparison of our model with Groups A, B, and C is as follows. Compared to A2, the accuracy improves by 11.61%, and the F1-score increases by 11.85% as C2 introduces the image modal to support text information. By adding cross-attention mechanisms, C2 enables the model to integrate information between text and images, thereby improving the accuracy of emotional analysis. Meanwhile, B4 only pairs the two modals together at a specific stage without considering the modals' correlation. Thus, C2 is 11.02% more accurate than B4, and F1-score is 8.58% better. Compared to B5, C2 introduces CPC, which helps to capture and utilize relevant information between text and images. Hence, the accuracy improves by 8.59%, and the F1-score increases by 3.82%. Compared to C1, C2 appends LoRA, which increases the model's efficiency by reducing the number of parameters. As a result, accuracy increases by 2.14%, and F1-score rises by 1.34%. These results indicate that the C2 has significantly improved emotional analysis tasks and confirmed the effectiveness and importance of cross-attention mechanisms, CPC, and LoRA in multimodal emotion analysis.

D. ABLATION EXPERIMENTS

To prove the necessity of every component in the proposed model, we carry out two sets of ablation experiments to achieve sentiment classification. The setup and results from the two ablation experiments are shown in the following paragraph.

1) LOSS FUNCTION ABLATION EXPERIMENT

Our model employs two training mechanisms: cross-entropy loss and contrastive learning loss. Cross-entropy loss guarantees the effectiveness of the sentiment classification. By contrast, the contrastive learning loss analyzes the dissimilarities between the combined modes of image and text features and the modes separately. The model utilizes two mechanisms to synthesize information from different modalities to improve accuracy. Then, we combine these two losses into a total loss function that optimizes the model's performance in multimodal sentiment analysis. We remove the two losses individually by conducting ablation experiments to evaluate their impact on the model's performance. The experimental results show the effects of different loss functions on the model's performance during the iterative process, revealing the importance of integrating other losses to improve model accuracy. The experimental results are shown in Figure 4, where cls_loss represents classification loss, cpc_loss represents contrastive learning loss, and loss represents total loss. The vertical axis indicates the value of the loss function, while the horizontal axis shows the number of iterations.

At the beginning of the iteration, all three losses plummet shockingly, while the total loss is higher than the others. The reason is that the total loss combines the sum of the

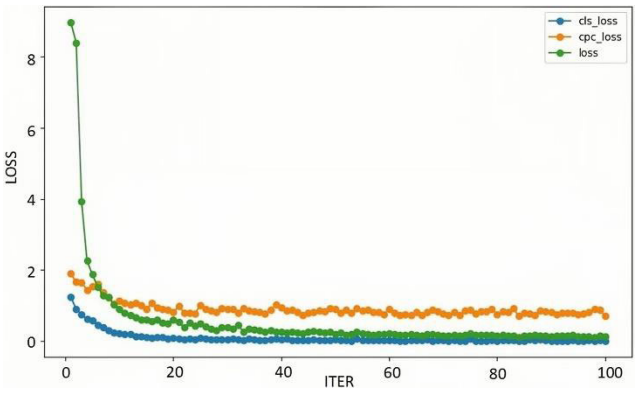

FIGURE 4. Results of the loss function ablation experiment.

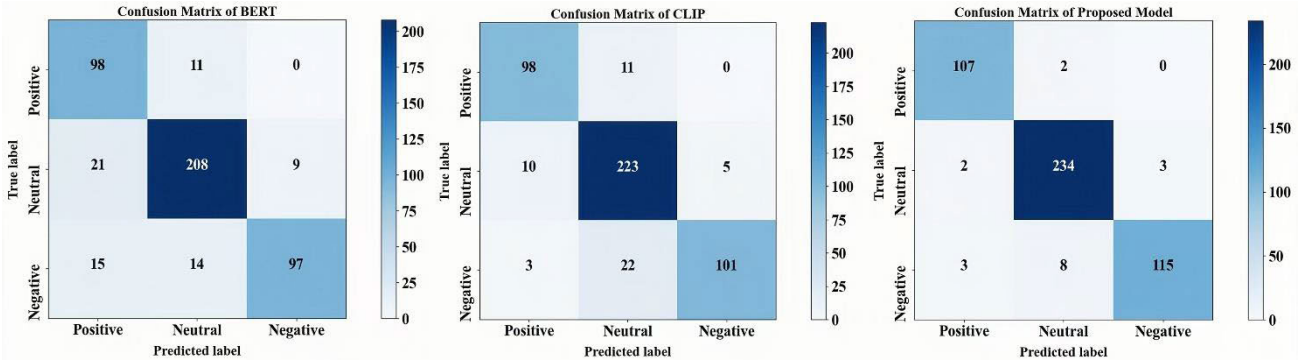

FIGURE 5. Confusion matrices for BERT, CLIP, and proposed model.

other two loss terms. As the number of iterations increases, the total loss tends to approach the cls_loss and cpc_loss after about seven iterations, showing a rapid response of the optimization. After 40 times iterations, the total loss becomes relatively stable, and the cpc_loss is always slightly higher than the cls_loss. The above situation indicates that the cls_loss contributes relatively little to the model optimization in the subsequent iterations, while the cpc_loss maintains some influence throughout the training process. The decreasing total loss trend shows consistent and stable performance improvement in the model, although its decline rate slows. In addition, the total loss curve in the figure is lower than the cpc_loss curve but higher than the cls_loss curve most of the time throughout the iteration, which indicates that the model can effectively acquire richer information from the data when the two losses are combined for training. Although the total loss is higher than the cpc_loss at some iteration points, the trend and level of the loss indicate that the model comprehensively considers cpc_loss and cls_loss to achieve better generalization performance.

2) MODALITY ABLATION EXPERIMENTS

We also conduct modality ablation experiments. Confusion matrices are used to analyze the errors of the proposed model. We perform confusion matrix analysis on BERT, CLIP, and the complete model, presenting the results in Figure 5. The full multimodal sentiment analysis model significantly improves performance in the experiments. The complete model proves to be the best. At the same time, the adaptation of CLIP outperforms BERT. Therefore, the positive effect of combining multiple fusion of information from various modalities is apparent. Meanwhile, the proposed multimodal model can fully utilize textual and image information to understand the context better and fill information gaps in each mode. In addition, effective fusion methods have played a significant role in enhancing the performance of sentiment analysis.

3) MISCLASSIFIED ANALYSIS

Misclassification is a concern in our study, especially the possible negative information transfer between modalities. The model is disrupted when the information from the text and visual elements is inconsistent or unclear. In this case, the sentiment tendency of the text may not match the sentiment conveyed by the image, leading to model confusion. For example, a positive image paired with negative text in ironic situations might cause misclassification. Therefore, further research on the interaction information transfer mechanism will help to mitigate this adverse effect and improve overall performance.

V. CONCLUSION AND OUTLOOK

A. CONCLUSION

In this paper, we propose a multimodal sentiment analysis architecture that uses contrastive learning and a cross-attention fusion network to study the sentiment of comments on governmental information across diversified media. We first analyze the demand for multimodal sentiment analysis in governmental social media. Then, we design cross-modal contrastive learning to study samples' intra and inter-class dynamics across modalities. Finally, we develop a feature fusion network based on cross-attention to integrate textual and visual features and a decision fusion method to assess sentiment analysis results across different modalities comprehensively.

Firstly, a new approach to effectively processing governmental information is proposed, combining contrastive learning and cross-entropy loss methods. This approach involves designing cross-modal contrastive learning to study the intra- and inter-class dynamics across modalities. The aim is to improve the efficiency of multi-task learning. Contrastive learning enhances the understanding of multimodal governmental information, particularly in processing text and image data. By learning the dynamics within classes across different modalities, the model improves its ability to align complex multimodal data and enhance its representational learning capabilities. The cross-entropy loss method improves the model's performance in sentiment analysis and other tasks, meeting the analysis requirements for comments in the governmental domain. Thus, the design proposed in this study effectively addresses the complexity and diversity

of multimodal sentiment analysis on comments related to governmental information.

Secondly, a feature fusion network is designated based on cross-attention to integrate textual and visual features. Our research on multimodal fusion methods indicates that interaction between different modalities is beneficial for the accuracy of sentiment analysis. Cross-attention provides the advantage of creating a more robust relationship between text and images. This technique aids the model in better understanding multimodal information. Using the feature fusion network can improve the accuracy and performance of sentiment analysis, ensuring the thorough realization of emotional expressions in governmental information. In addition, we adopt a decision fusion approach, considering the sentiment analysis results from various modalities collectively. This integration balances the contributions of different modalities, ensuring more accurate and comprehensive sentiment analysis results. Therefore, our design enhances the performance and usability of multimodal sentiment analysis in governmental information comments.

Finally, our novel multimodal sentiment analysis model demonstrates superior performance with an accuracy rate of 96.80% and an F1-score of 94.23%. Compared with the best recognition achieved through unimodal, the proposed method improves recognition by 11.61% and 11.85%. The model could assist the government in emotional evolution analysis, detection of public opinion, and online public opinion guidance. The government may also need to respond to complex social feedback, improve governance efficiency, and enhance public participation. This efficient multimodal sentiment analysis framework offers a new perspective in governmental decision-making, helping address information society's challenges.

B. LIMITATIONS AND FUTURE RESEARCH

Although our proposed model shows high accuracy in recognition, it does not exhibit overfitting to any particular modality. Because we consider both modalities' information at the feature extraction stage, and the fusion strategy also effectively balances the contribution of text and image features. Furthermore, we can explore more sophisticated multimodal fusion networks to maximize sentiment recognition performance by fully leveraging the potential of each modality. In future research, we plan to conduct experiments on various social media platforms, such as Twitter, to validate our findings further and ensure the robustness of our model. These improvements and expansions will help address the challenges of sentiment analysis in different social media contexts more effectively.

REFERENCES

[1] G. Mu, J. Li, Z. Liao, and Z. Yang, "An enhanced IHHO-LSTM model for predicting online public opinion trends in public health emergencies," *Sage Open*, vol. 14, no. 2, Apr. 2024, Art. no. 21582440241257681, doi: 10.1177/21582440241257681.

[2] G. Mu, J. Li, X. Li, C. Chen, X. Ju, and J. Dai, "An enhanced IDBO-CNN-BiLSTM model for sentiment analysis of natural disaster tweets," *Biomimetics*, vol. 9, no. 9, p. 533, Sep. 2024, doi: 10.3390/biomimetics9090533.

[3] Q. Xianjun, C. Minghong, and L. Xiaoli, "User acceptance model of government microblog and its empirical study," *Proc. Comput. Sci.*, vol. 162, pp. 940–945, Jan. 2019, doi: 10.1016/j.procs.2019.12.071.

[4] J. Wang, X. Wang, and L. Fu, "Evolutionary game model of public opinion information propagation in online social networks," *IEEE Access*, vol. 8, pp. 127732–127747, 2020, doi: 10.1109/ACCESS.2020.3006150.

[5] H.-H. Nguyen and M.-T. Nguyen, "Emotion-cause pair extraction as question answering," in *Proc. 15th Int. Conf. Agents Artif. Intell.*, 2023, pp. 988–995, doi: 10.5220/0011883100003393.

[6] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, pp. 306–325, Jul. 2023, doi: 10.1016/j.inffus.2023.02.028.

[7] J. Khan, N. Ahmad, S. Khalid, F. Ali, and Y. Lee, "Sentiment and context-aware hybrid DNN with attention for text sentiment classification," *IEEE Access*, vol. 11, pp. 28162–28179, 2023, doi: 10.1109/ACCESS.2023.3259107.

[8] X. Chen, W. Zhang, X. Xu, and W. Cao, "A public and large-scale expert information fusion method and its application: Mining public opinion via sentiment analysis and measuring public dynamic reliability," *Inf. Fusion*, vol. 78, pp. 71–85, Feb. 2022, doi: 10.1016/j.inffus.2021.09.015.

[9] C. Cai, Y. He, L. Sun, Z. Lian, B. Liu, J. Tao, M. Xu, and K. Wang, "Multimodal sentiment analysis based on recurrent neural network and multimodal attention," in *Proc. 2nd Multimodal Sentiment Anal. Challenge*, Oct. 2021, pp. 61–67, doi: 10.1145/3475957.3484454.

[10] T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez, "A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102040, doi: 10.1016/j.inffus.2023.102040.

[11] W. Shi, J. Zhang, and S. He, "Understanding public opinions on Chinese short video platform by multimodal sentiment analysis using deep learning-based techniques," *Kybernetes*, Sep. 2023, doi: 10.1108/k-04-2023-0723.

[12] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 164–172, doi: 10.1609/aaai.v34i01.5347.

[13] H. Wang, W. Zhang, and X. Ma, "Contrastive and adversarial regularized multi-level representation learning for incomplete multi-view clustering," *Neural Netw.*, vol. 172, Apr. 2024, Art. no. 106102, doi: 10.1016/j.neunet.2024.106102.

[14] J. Li and X. Wu, "Simple framework for the contrastive learning of visual representations-based data-driven tight frame for seismic denoising and interpolation," *Geophysics*, vol. 87, no. 5, pp. 467–480, Aug. 2022, doi: 10.1190/geo2021-0590.1.

[15] T. Zhao, L.-A. Meng, and D. Song, "Multimodal aspect-based sentiment analysis: A survey of tasks, methods, challenges and future directions," *Inf. Fusion*, vol. 112, Dec. 2024, Art. no. 102552, doi: 10.1016/j.inffus.2024.102552.

[16] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," 2018, *arXiv:1805.11730*.

[17] H. Cheng, Z. Yang, X. Zhang, and Y. Yang, "Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3149–3163, Oct. 2023, doi: 10.1109/TAFFC.2023.3265653.

[18] L. Feng, L.-Y. Liu, S.-L. Liu, J. Zhou, H.-Q. Yang, and J. Yang, "Multimodal speech emotion recognition based on multi-scale MFCCs and multi-view attention mechanism," *Multimedia Tools Appl.*, vol. 82, no. 19, pp. 28917–28935, Mar. 2023, doi: 10.1007/s11042-023-14600-0.

[19] Y. Li, Y. Li, S. Zhang, G. Liu, Y. Chen, R. Shang, and L. Jiao, "An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency," *Knowl.-Based Syst.*, vol. 287, Mar. 2024, Art. no. 111457, doi: 10.1016/j.knosys.2024.111457.

[20] M. Huang, H. Xie, Y. Rao, Y. Liu, L. K. M. Poon, and F. L. Wang, "Lexicon-based sentiment convolutional neural networks for online review analysis," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1337–1348, Jul. 2022, doi: 10.1109/TAFFC.2020.2997769.

[21] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, and X. Huang, "Sentiment-aware multimodal pre-training for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 258, Dec. 2022, Art. no. 110021, doi: 10.1016/j.knosys.2022.110021.

[22] Z. Li, Q. Guo, Y. Pan, W. Ding, J. Yu, Y. Zhang, W. Liu, H. Chen, H. Wang, and Y. Xie, "Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101891, doi: 10.1016/j.inffus.2023.101891.

[23] R. Gupta, J. Kumar, and H. Agrawal, "A statistical approach for sarcasm detection using Twitter data," in *Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2020, pp. 633–638, doi: 10.1109/iciccs48265.2020.9120917.

[24] G. Mu, N. Gao, Y. Wang, and L. Dai, "A stock price prediction model based on investor sentiment and optimized deep learning," *IEEE Access*, vol. 11, pp. 51353–51367, 2023, doi: 10.1109/ACCESS.2023.3278790.

[25] K.-X. Han, W. Chien, C.-C. Chiu, and Y.-T. Cheng, "Application of support vector machine (SVM) in the sentiment analysis of Twitter DataSet," *Appl. Sci.*, vol. 10, no. 3, p. 1125, Feb. 2020, doi: 10.3390/app10031125.

[26] R. Piryani, B. Piryani, V. K. Singh, and D. Pinto, "Sentiment analysis in nepali: Exploring machine learning and lexicon-based approaches," *J. Intell. Fuzzy Syst.*, vol. 39, no. 2, pp. 2201–2212, Aug. 2020, doi: 10.3233/jifs-179884.

[27] J. Kazmaier and J. H. van Vuuren, "The power of ensemble learning in sentiment analysis," *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115819, doi: 10.1016/j.eswa.2021.115819.

[28] X. Fan, A. Runa, Z. Pei, and M. Jiang, "An improved convolutional neural network for text classification," *J. Phys., Conf.*, vol. 2066, no. 1, Nov. 2021, Art. no. 012091.

[29] J. Z. Zhang, P. R. Srivastava, and P. Eachempati, "Evaluating the effectiveness of drones in emergency situations: A hybrid multi-criteria approach," *Ind. Manage. Data Syst.*, vol. 123, no. 1, pp. 302–323, Nov. 2021, doi: 10.1108/imds-01-2021-0064.

[30] H. Chen, M. Wang, and Z. Zhang, "Research on rural landscape preference based on TikTok short video content and user comments," *Int. J. Environ. Res. Public Health*, vol. 19, no. 16, p. 10115, Aug. 2022, doi: 10.3390/ijerph191610115.

[31] H. Lei, H. Wang, L. Wang, Y. Dong, J. Cheng, and K. Cai, "An analysis of the evolution of online public opinion on public health emergencies by combining CNN-BiLSTM + attention and LDA," *J. Comput. Commun.*, vol. 11, no. 4, pp. 190–199, 2023, doi: 10.4236/jcc.2023.114009.

[32] A. R. Rahmanti, C.-H. Chien, A. A. Nursetyo, A. Husnayain, B. S. Wiratama, A. Fuad, H.-C. Yang, and Y.-C.-J. Li, "Social media sentiment analysis to monitor the performance of vaccination coverage during the early phase of the national COVID-19 vaccine rollout," *Comput. Methods Programs Biomed.*, vol. 221, Jun. 2022, Art. no. 106838, doi: 10.1016/j.cmpb.2022.106838.

[33] R. Lerouge, M. D. Lema, and M. Arnaboldi, "The role played by government communication on the level of public fear in social media: An investigation into the covid-19 crisis in Italy," *Government Inf. Quart.*, vol. 40, no. 2, Apr. 2023, Art. no. 101798, doi: 10.1016/j.giq.2022.101798.

[34] M. Thomson, H. Murfi, and G. Ardaneswari, "BERT-based hybrid deep learning with text augmentation for sentiment analysis of Indonesian hotel reviews," in *Proc. 12th Int. Conf. Data Sci., Technol. Appl.*, 2023, pp. 468–473, doi: 10.5220/0012127400003541.

[35] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–19, Jan. 2023, doi: 10.1145/3517139.

[36] S. Zou, X. Huang, X. Shen, and H. Liu, "Improving multimodal fusion with main modal transformer for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 258, Dec. 2022, Art. no. 109978, doi: 10.1016/j.knosys.2022.109978.

[37] S. Zhang, M. Chen, J. Chen, Y.-F. Li, Y. Wu, M. Li, and C. Zhu, "Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107340, doi: 10.1016/j.knosys.2021.107340.

[38] W. Jin, B. Zhao, L. Zhang, C. Liu, and H. Yu, "Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis," *Inf. Process. Manage.*, vol. 60, no. 3, May 2023, Art. no. 103260, doi: 10.1016/j.ipm.2022.103260.

[39] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–12, doi: 10.18653/v1/d17-1115.

[40] M. Chen, S. Wang, P. P. Liang, T. Baltrusaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 163–171, doi: 10.1145/3136755.3136801.

[41] V. Pérez-Rosas, R. Mihalcea, and L. P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Aug. 2013, pp. 973–982.

[42] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. 18th ACM Int. Conf. Multimodal Interaction*, Oct. 2016, pp. 284–288, doi: 10.1145/2993148.2993176.

[43] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 949–954, doi: 10.1109/ICME.2017.8019301.

[44] M. S. Hussain, R. A. Calvo, and P. A. Pour, "Hybrid fusion approach for detecting affects from multichannel physiology," in *Affective Computing and Intelligent Interaction* (Lecture Notes in Computer Science), Berlin, Germany: Springer, 2011, pp. 568–577, doi: 10.1007/978-3-642-24600-5_60.

[45] J. Zhang and L. Jiang, "Research on irony recognition of travel reviews based on multimodal deep learning," *Inf. Studies, Theory Appl.*, vol. 7, pp. 158–164, Jan. 2022, doi: 10.16353/j.cnki.1000-7490.2022.07.022.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[47] 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 *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558, doi: 10.18653/v1/p19-1656.

[48] D. Hazarika, R. Zimmermann, and S. Poria, "MISA," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131, doi: 10.1145/3394171.3413678.

[49] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018, doi: 10.1016/j.knosys.2018.07.041.

[50] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 873–883, doi: 10.18653/v1/p17-1081.

[51] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 3454–3466, doi: 10.18653/v1/d18-1382.

GUANGYU MU received the bachelor's degree in computer science from Jilin Engineering University, in 1994, and the master's and Ph.D. degrees in computer science from Jilin University, in 2004 and 2011, respectively. She is currently a Professor with the School of Management Science and Information Engineering and the Director of the Institute of Economic Information Management, Jilin University of Finance and Economics. She is a Peer Reviewer of China National Social Science Fund Projects and an Expert on Humanities and Social Science Fund Projects of the Ministry of Education of China and the Outstanding Contribution of Jilin Province. Her research interests include information resource management, data mining, and financial engineering.

CHUANZHI CHEN received the Literature degree, in 2023. He is currently pursuing the master's degree in library and information science with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. His current research interests include multimodal sentiment analysis and multimodal disinformation detection. He won the Third-Class Scholarship.

XIURONG LI received the bachelor's degree in computer science from Jilin Engineering University, in 1994, and the master's degree in computer science from Beijing University of Technology, in 2005. She is currently an Associate Professor with Beijing University of Technology. Her research interests include cryptography and information security.

JIAXUE LI received the bachelor's degree in management, in 2022. He is currently pursuing the master's degree in library and information studies with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. His research interests include online public opinion and information resource management.

XIAOQING JU received the Engineering degree, in 2023. She is currently pursuing the master's degree in electronic information with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. Her current research interests include multimodal fusion and false information detection. She won the Third-Class Scholarship.

JIAXIU DAI received the Engineering degree, in 2024. She is currently pursuing the master's degree in electronic information with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. Her current research interest includes fake news detection. She won the Third-Class Scholarship.