Fine-Grained Sentiment Analysis of Public Opinion Videos Based on Conformer and Multi-Layered Interaction Attention

Chaolong Liu, Zhengguang Gao, Lihong Zhang*
Research Center for Network Public Opinion Governance, China People's Police University, Langfang, China
**Corresponding Author.*

Abstract: With the rapid development of internet technologies and the widespread adoption of smart devices, social media platforms have become significant channels for information dissemination and public sentiment expression. In particular, new media formats such as short videos have shown a substantial impact on public opinion guidance and emotional transmission, making sentiment analysis of short video content highly meaningful. However, existing research has limitations in modality interactions, often employing weighted summation or self-attention mechanisms for deep fusion of extracted features. These approaches fail to fully account for the complex local dependencies and hierarchical structures among modalities. To address these issues, this paper proposes a fine-grained sentiment analysis model for public opinion videos based on Conformer and multi-layered interaction attention mechanisms, termed DW-MIACon. The model first utilizes DeBERTa, CLIP, and Wav2Vec models to extract features from text, images, and audio, respectively. Subsequently, the extracted multimodal features are fused using a Dynamic Weighted Multi-layered Interaction Attention (DW-MIA) mechanism, generating rich fusion feature representations. Finally, a Conformer model is employed to deeply integrate the fused features, capturing complex interactions and local dependencies between modalities. Experimental results demonstrate that the proposed model significantly outperforms existing approaches in multimodal sentiment recognition tasks, notably enhancing the accuracy of fine-grained sentiment classification and the ability to identify subtle emotional nuances.

Keywords: Public Opinion; Conformer; Multi-layered Interaction Attention; Fine-grained; Sentiment Analysis

1. Introduction

In today's society, characterized by the rapid advancement of internet technologies and the widespread adoption of smart devices, social media platforms have gradually become key channels for information acquisition, opinion expression, and interactive communication. According to the latest statistics released by the China Internet Network Information Center (CNNIC), as of August 2024, China's internet penetration rate has reached 78.0%, with nearly 1.1 billion internet users. This vast user base not only provides social media platforms with rich data resources but also offers a broad research domain for public opinion studies. Research has shown that short videos, due to their ease of dissemination, real-time updates, and high interactivity, have significant advantages in information propagation and public opinion guidance. Short video applications, represented by platforms like Douyin and Kuaishou, exhibit explosive growth in the speed and influence of online public opinion dissemination. In the post-truth era, public emotions often surpass objective facts, becoming a powerful force that shapes online public opinion[1]. However, this phenomenon is also accompanied by the proliferation of misinformation and the spread of negative public sentiment, posing potential threats to social stability. Therefore, effectively monitoring, deeply analyzing, and appropriately responding to short video content have become critical issues urgently needing exploration within the academic community.

Sentiment analysis aims to identify the underlying emotional tendencies in data by analyzing various modalities such as text,

images, and audio. Traditional sentiment analysis research mainly focuses on coarse-grained sentiment recognition at the macro level, such as classifying sentiments into positive, negative, or neutral categories. In contrast, fine-grained sentiment analysis provides more detailed sentiment labels, offering a more precise reflection of users' emotional states. This is particularly significant for understanding and analyzing user behavior patterns on social media.

The core challenge of multimodal sentiment analysis lies in effectively integrating data from different modalities to accurately identify emotional information. Multimodal fusion methods can be broadly categorized into three types: early fusion, late fusion, and hybrid fusion. Early fusion involves the direct combination and input-level fusion of low-level features, which has been proven effective in predicting multi-label sentiments. Late fusion, on the other hand, integrates features from each modality during the decision phase, preserving the independence of modalities but often overlooking the interaction details between them. Hybrid fusion employs a hierarchical approach to fuse features between modalities at different stages and is one of the current research focuses. Williams et al. [2] combined low-level feature fusion with bidirectional long short-term memory networks (Bi-LSTM), demonstrating the effectiveness of early fusion techniques in multi-label prediction. Zadeh et al. [3] proposed the Memory Fusion Network (MFN), a neural network architecture for multi-view learning that interprets interactions across different modalities and sequentially models multimodal features at the same time step. In recent years, self-attention mechanisms, such as those used in BERT and CLIP, have been widely applied in multimodal sentiment analysis, showing excellent performance in feature extraction. The MARN model[4]utilizes a multi-layer attention mechanism to explore cross-modal emotional contexts within time steps, storing these contexts in hybrid memory blocks, thus enhancing the understanding of cross-modal associations and improving the use of cross-modal information for sentiment analysis.However, existing methods often adopt weighted summation or self-attention mechanisms during the modality fusion stage, frequently neglecting the complex local dependencies and hierarchical structures among modalities. To enhance modality interaction, many studies employ strategies that independently extract unimodal features and directly fuse them, failing to fully explore the complex relationships between modalities. Han et al.[5] introduced the concept of bimodal fusion by calculating modality-related increments and modality difference increments separately, training two components to learn their probability distributions. Williams et al.[2] proposed a sequence learning approach based on input-level feature fusion and bidirectional long short-term memory (BLSTM) deep neural networks (DNNs), where audio, video, and text modalities are fused at the input level for emotion recognition. While these methods improve fusion effectiveness to some extent, they still fall short in capturing the comprehensive interactions among the three modalities.

To address these challenges, this paper proposes a fine-grained sentiment analysis model for public opinion videos based on Conformer and Multi-layered Interaction Attention, termed DW-MIACon. Initially, the DeBERTa, CLIP, and Wav2Vec models are employed to extract features from text, image, and audio data, respectively. These features are fused using a Dynamic Weighted Multi-layered Interaction Attention (DW-MIA) mechanism, which generates bimodal fusion features that are further integrated with the third modality, enabling a deeper exploration of the interrelationships and complementarity among different modalities. This process generates richer modality feature representations. Finally, the Conformer model is utilized to perform deep fusion of these multimodal features. By combining convolution and self-attention mechanisms, the Conformer model effectively captures local dependencies among modalities, enhancing the depth and efficacy of the fusion process. Through this comprehensive approach, a high-dimensional, multimodal feature fusion vector is obtained for sentiment analysis.

2. Related Work

Multimodal sentiment analysis aims to utilize a combination of various data modalities, such as text, audio, images, and videos, to perform sentiment recognition and analysis. Compared to unimodal

sentiment analysis, multimodal sentiment analysis can leverage the features of different modalities to capture richer emotional cues, effectively compensating for the limitations of single modalities in terms of information loss and misinterpretation. By integrating multimodal information, this approach provides a more comprehensive and precise understanding of the complexity and diversity of emotions, making it widely applicable in fields like sentiment analysis and social media analytics.

The choice of modality fusion strategies is crucial in multimodal sentiment analysis, as it directly impacts the integration of information from different modalities and the accuracy of the final sentiment recognition. Early fusion strategies primarily include early fusion and late fusion methods. Morency et al.[6]conducted pioneering work in the tri-modal sentiment analysis task by automatically extracting features from text, visual, and audio data, concatenating them, and feeding the combined features into a Hidden Markov Model (HMM) for classification. This work provided preliminary validation of the feasibility of multimodal sentiment analysis. Yu et al.[7] further utilized Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) to extract features from text and visual data, optimizing the process through averaging or weighted fusion strategies. However, these methods have limitations in terms of modality interaction, failing to fully explore the deep associations and interdependencies between modalities. To address these challenges, researchers have proposed hybrid fusion strategies. Hybrid fusion not only considers the synergistic relationships between modalities but also dynamically adjusts the weights of each modality at different stages. Although hybrid fusion strategies significantly enhance model performance, their design and training process are complex, requiring more computational resources and meticulous parameter tuning.

With the rapid development of deep learning and attention mechanisms, their application in multimodal sentiment analysis has become increasingly prevalent. Tsai et al.[8] introduced the MulT model, which employs bidirectional cross-modal attention mechanisms to enable effective interactions between multimodal sequences, improving the accuracy of sentiment recognition. Yang et al.[9] developed the Cross-Modal BERT (CM-BERT), which integrates the interactions between text and audio modalities into the pre-trained BERT model, yielding enhanced feature representations. Feng Cheng et al.[10] proposed a multimodal sentiment analysis model based on top-down mask generation and stacked Transformers. This model generates feature masks through a mask generation module and applies them to other modalities, effectively exploring the relationships and complementarities among different modalities. Junjie Wu et al.[11] introduced a bimodal sentiment computation model, which uses Multilayer Perceptron (MLP) and Bidirectional Long Short-Term Memory (BiLSTM) networks for feature extraction. The extracted features are then fused using MLP and self-attention mechanisms.

3. Methods

The proposed DW-MIACon framework for fine-grained sentiment analysis of public opinion videos, as illustrated in **Figure 1.**, consists of four main modules: feature extraction, modality interaction, modality fusion, and sentiment classification. In the feature extraction module, the DeBERTa, CLIP, and Wav2Vec models are employed to extract features from text, image, and audio data, respectively. The modality interaction module employs a Dynamic Weighted Multi-layered Interaction Attention (DW-MIA) mechanism, which performs multi-layered interaction and fusion of the extracted features. This mechanism dynamically weights and adjusts the features of different modalities, effectively enhancing the synergistic effects between them. In the modality fusion module, a Conformer model is used to deeply fuse the interacted features. The Conformer model combines convolutional and self-attention mechanisms to capture local dependencies between modalities, comprehensively considering both detailed and overall information of multimodal features, thereby improving the depth and effectiveness of feature fusion. Finally, these deeply fused

features are fed into the sentiment classification module, which performs sentiment analysis to generate fine-grained sentiment classification results for public opinion videos.

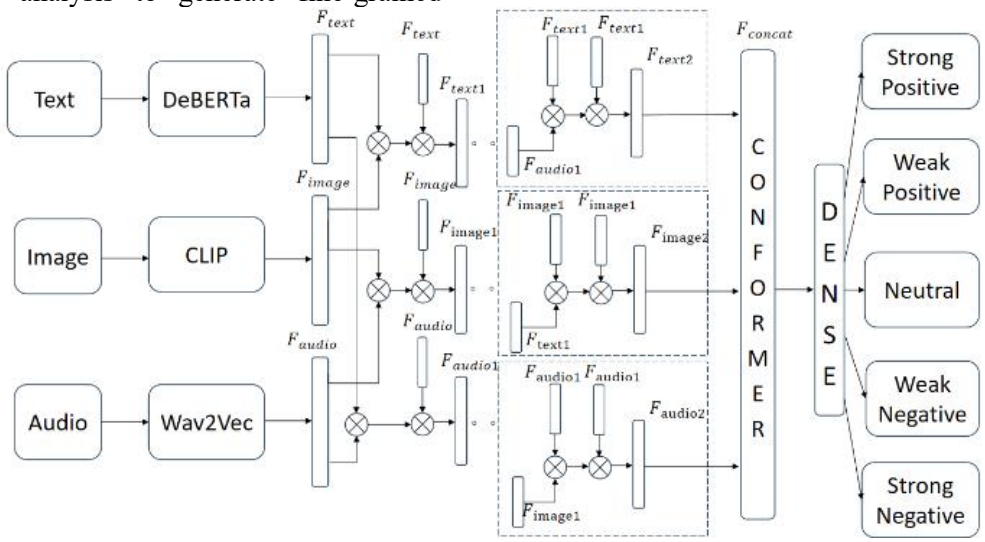

Figure 1. Framework of Fine-Grained Sentiment Analysis Model for Public Opinion Videos

3.1 Feature Extraction Module

The feature extraction module is fundamental to the proposed fine-grained sentiment analysis model for public opinion videos, responsible for extracting high-quality modality features from text, image, and audio data. This module provides rich feature representations for subsequent multimodal interaction and fusion by utilizing the DeBERTa, CLIP, and Wav2Vec models to process text, image, and audio data, respectively, ensuring precise and comprehensive representation of the data features of each modality.

For text data, this study employs the DeBERTa model as the feature extraction tool. DeBERTa, a language model based on the Transformer architecture, introduces disentangled attention mechanisms and an enhanced mask decoder, significantly improving the performance of masked language models[12]. The model separates context information into content and positional aspects, enabling more precise capture of text semantics, particularly enhancing its ability to represent complex semantic and syntactic structures.

Let the input text sequence be $T = \{t_1, t_2, \ldots, t_n\}$, The feature representation of the text, F_{text} , is obtained using the DeBERTa model, as illustrated in Formula (1).

$$F_{\text{text}} = \text{DeBERTa}(T) \quad (1)$$

The CLIP model, a multimodal pre-training model based on contrastive learning, maps images and text into the same semantic space, effectively extracting semantic features of images[13]. The Vision Transformer (ViT) module of CLIP extracts features from images, enabling comparative learning between image and text features during multimodal interaction, thus enhancing the alignment between the modalities. Consequently, this study selects the CLIP model for image feature extraction. Given an input image I , the feature representation F_{image} is extracted using CLIP's ViT module, as illustrated in Formula (2).

$$\text{F}_{\text{image}} = \text{ViT(I)} \quad (2)$$

Here, the image I is divided into several patches, processed through patch embedding and positional encoding, and finally transformed into the global semantic representation F_{image} using the Transformer layers.

Audio feature extraction is performed using the Wav2Vec model[14], a self-supervised learning-based speech feature extraction model. Wav2Vec captures high-dimensional speech features from large amounts of unlabeled audio data through self-supervised learning, effectively capturing temporal and spectral characteristics of the audio. These features provide precise acoustic information for sentiment analysis, particularly excelling in capturing emotional variations and tonal expressions.

Thus, Wav2Vec is chosen as the audio feature extraction model. Given an input audio signal A, the feature representation F_{audio} is extracted, as illustrated in Formula (3).

$$\mathrm{F_{audio}} = \mathrm{Wav2Vec(A)} \quad (3)$$

The audio signal first undergoes preliminary processing through the Feature Extractor, followed by high-dimensional feature representation learning via the Encoder, ultimately yielding the audio features F_{audio} for subsequent interactions.

3.2 Dynamic Weighted Multi-layered Interaction Attention Mechanism

In multimodal sentiment analysis tasks, the information embedded within different modalities such as text, image, and audio exhibits significant heterogeneity and complementarity. Each modality uniquely contributes to emotional expression, and a single modality often fails to capture the subtle nuances of sentiment comprehensively. Thus, effectively enhancing the interactions between modalities to extract more comprehensive and profound emotional information remains a key challenge in multimodal sentiment analysis. Additionally, the importance of each modality in expressing sentiment can vary depending on the context. For instance, audio may convey more emotional information than images in certain scenarios, whereas text might serve as the primary carrier of sentiment in others. Consequently, static feature fusion methods may not fully exploit the potential of each modality.

To address these issues, this paper proposes a Dynamic Weighted Multi-layered Interaction Attention mechanism (DW-MIA). The DW-MIA mechanism is composed of multiple interaction modules, each incorporating a dynamic weighting strategy that provides adaptable modal weight allocation for each layer of the interaction attention modules. This design enables layer-by-layer interaction and fusion of modality features, facilitating smoother information flow between modalities while adaptively adjusting the influence of each modality. As a result, it enhances the exploration of hidden inter-modal relationships, focuses more on critical modalities, and improves the ability to capture subtle emotional information.

Taking text features as an example, the interaction between text and image modalities is first achieved through a Cross Attention Mechanism. This mechanism calculates the interaction weights between audio and image modalities, enabling weighted feature fusion[15].

$$\mathrm{image\_text\_attention} = \mathrm{Attention}(\mathrm{query} = \mathrm{F_{text}}, \mathrm{key} = \mathrm{F_{image}}, \mathrm{value} = \mathrm{F_{image}}) \quad (4)$$

After the initial cross-attention calculations, the model introduces a dynamic weight computation layer. This layer calculates dynamic weights w for each modality through a fully connected layer and normalizes them using a Softmax activation function. The weight computation is described by the following formula (5):

$$\mathrm{w} = \mathrm{Softmax(Dense(fusion\ x))} \quad (5)$$

where $\mathrm{fusion\ x}$ refers to the initial fused features (i.e., the concatenation of image, audio, and text features). The dynamic weight w is split into three components: w_v、w_t、w_a corresponding to the image, text, and audio modalities, respectively.

The dynamically computed weights are then used to perform weighted fusion of each modality's features. For example, text features are weighted through image-text cross-attention and dynamic weights, as shown in the following equations (6-7):

$$\mathrm{F_{text1}} = \mathrm{F_{text}} + \mathrm{w_t} * \mathrm{text\_image\_attention} \quad (6)$$

$$\mathrm{F_{text2}} = \mathrm{F_{text1}} + \mathrm{w_t} * \mathrm{text\_aud_{attention}} \quad (7)$$

where w_t is the dynamic weight associated with the text modality, F_{text1} represents the text features after bi-modal interaction, and F_{text2} denotes the text features after tri-modal interaction. The symbol $*$ denotes the multiplication operation. The $\mathrm{text\_image\_attention}$ refers to the text-image bimodal fusion features, while $\mathrm{text\_aud\_attention}$ refers to the text-audio bimodal fusion features.

Similarly, audio and image features undergo interaction with corresponding dynamic weights. The dynamic weighted multi-layered interaction attention mechanism enables multi-level interactions and deep fusion of features from different modalities, while dynamically adjusting each modality's impact on sentiment recognition. This adaptive feature adjustment strategy focuses on the most representative

modality features based on the actual sentiment expression needs, thereby significantly enhancing the accuracy and robustness of sentiment recognition. Additionally, the multi-layered interaction design allows for a more comprehensive exploration of the subtle relationships between modalities, improving the overall expressive capability of the sentiment analysis model.

3.3 The Conformer-Based Multimodal Feature Fusion

The Transformer architecture has demonstrated outstanding performance in various sequence tasks. However, its self-attention mechanism has limitations in capturing local information, which can lead to performance degradation in tasks such as speech processing and long text sequence analysis. In contrast, Convolutional Neural Networks (CNNs) excel at capturing local features. To address these challenges, the Conformer model was introduced, combining the strengths of convolution and self-attention mechanisms. By embedding convolutional modules within the Transformer architecture, the Conformer not only enhances the ability to capture global information but also effectively addresses the shortcomings of traditional self-attention mechanisms in handling local dependencies.

The core architecture of the Conformer model consists of four main components: a Feed-Forward Neural Network (FFN) module, a Multi-Head Self-Attention (MHSA) module, a Convolutional module, and a Layer Normalization module. These modules are arranged in a specific sequence, forming a complete Conformer unit, as illustrated in **Figure 2.**

In the feature fusion module, the text features $F_{text2} \in R^{T_t \times d_t}$, image features $F_{image2} \in R^{T_i \times d_i}$, and audio features $F_{audio2} \in R^{T_a \times d_a}$, extracted from the feature interaction module, are first merged into a unified fusion feature matrix $\boldsymbol{F_{concat}}$. This fusion feature matrix is then fed into multiple Conformer units for deep feature interaction and information extraction:

$$F_{concat} = \left[F_{text2}; F_{image2}; F_{audio2}\right] \in R^{(T_t+T_i+T_a)\times d} \quad (8)$$

where T_t, T_i, T_a denote the sequence lengths of the text, image, and audio features, respectively, and d_t, d_i, d_a represent the dimensions of the text, image, and audio features.

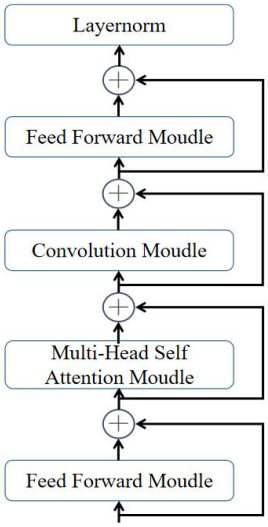

Figure 2. Conformer Structure

The input features are first processed by the Feed-Forward Neural Network (FFN) module. In the Conformer, the FFN module consists of two linear transformations with a nonlinear activation function designed to enhance the model's stability and nonlinear representation capabilities. To further improve performance, the Conformer introduces an FFN module both before and after the self-attention module. The processing steps are as follows equations (9-11):

$$\mathrm{F_{ffn1}} = \mathrm{FFN_1(F)} = \mathrm{RELU(FW_1 + b_1)} \quad (9)$$

$$F_{ffn2} = FFN_2\left(F_{ffn1}\right) = F_{ffn1}W_2 + b_2 \quad (10)$$

Finally, a residual connection is added:

$$F_{ffn} = F + F_{ffn2} \quad (11)$$

FFN_1 and FFN_2 are feed-forward neural network layers, including linear transformations and activation functions. The residual connection $(+)$ enhances the model's stability and expressive capacity.

Next, the Multi-Head Self-Attention (MHSA) module is responsible for capturing long-range dependencies between different modalities. This module computes attention weights over various parts of the input sequence, enabling dynamic information weighting. By employing the multi-head mechanism, the Conformer learns attention distributions from multiple representation spaces, enriching feature representation. The calculation process is represented by the following equations (12-

15):

$$Q, K, V = F_{ffn}W_Q, F_{ffn}W_K, F_{ffn}W_V \quad (12)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

The output of the multi-head self-attention is:

$$F_{attn} = Concat(head_1, \ldots, head_h)W_o \quad (14)$$

Adding residual connection and normalization:

$$F_{attnout} = \text{LayerNorm}(F_{ffn} + F_{attn}) \quad (15)$$

Here, Q、K、V are the query, key, and value matrices obtained through linear transformations, and d_k is the dimension of the key vectors. $head_i$ is the output of the i attention head, and W_O is the output linear transformation matrix.

The convolution module is a key feature that distinguishes the Conformer from the traditional Transformer architecture, as it captures local information within features through 1D convolution. The Conformer employs Depthwise Separable Convolution to effectively capture local temporal features while reducing the number of parameters. The main steps of the convolution module include layer normalization, 1D pointwise convolution, Gated Linear Unit (GLU) activation, and depthwise convolution. The calculation process is represented by equations (16-17):

$$F_{conv} = \text{GLU}(F_{attnout} * W_{conv1}) * W_{conv2} \quad (16)$$

where GLU denotes the Gated Linear Unit, $*$ indicates the product operation, and W_{conv1} and W_{conv2} are the convolution kernel matrices.

A residual connection and normalization are then added:

$$F_c = \text{LayerNorm}(F_{conv} + F_{attnout}) \quad (17)$$

Subsequently, F_c is input into the second feed-forward module, following the same calculations as previously shown, resulting in F_{ffn}'.

Finally, the output features from the second feed-forward module, F_{ffn}' , undergo Layer Normalization to ensure consistent data distribution across different layers, aiding in stabilizing the training process. The calculation of Layer Normalization is represented by equation (18):

$$F_{out} = \text{LayerNorm}(F_{ffn}') \quad (18)$$

After processing through multiple Conformer blocks, the deeply fused features F_{out} are obtained.

Through its multi-level combination of feed-forward, self-attention, and convolution modules, the Conformer achieves joint modeling of global and local features. The fused features F_{out} not only enhance the interaction relationships between modalities but also preserve the key information of each modality's features, providing a rich input for subsequent sentiment classification tasks.

3.4 Sentiment Classification Module

The primary goal of the sentiment classification module is to conduct the final sentiment analysis using the comprehensive features extracted from different modalities in the video (text, image, and audio) and to produce fine-grained sentiment classification results. First, the deeply fused multimodal features from the Conformer model are fed into the classification network. The core component of this network is a fully connected (Dense) layer, which transforms the high-dimensional fused features F_{out} into a fixed-dimensional vector h corresponding to different sentiment categories. The operation of the Dense layer can be expressed as Equation (19):

$$h = W \cdot F_{out} + b \quad (19)$$

where W is the weight matrix and b is the bias term.

To enhance classification accuracy, a ReLU activation function and a Dropout layer are added after the Dense layer. The ReLU activation function is defined as Equation (20):

$$\text{ReLU} = \max(0, x) \quad (20)$$

The processed feature vector is then passed through a Softmax layer to compute the probability distribution for each sentiment category. The Softmax function is given by Equation (21):

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (21)$$

where z_i represents the score for the iii-th category, and C is the total number of sentiment categories. The Softmax function converts these scores into a probability distribution, reflecting the likelihood of each sentiment category[16].

Finally, the model selects the sentiment category with the highest probability as the sentiment label of the video. This step is

defined by Equation (22):

$$\text{Emotion} = \arg\max_i \left(\text{Softmax}(z_i)\right) \quad (22)$$

This step identifies the most likely sentiment category by comparing the probabilities of each sentiment class, generating the sentiment classification result for the video.

4. Experiment

4.1 Dataset

As public attention to police enforcement and related policies continues to grow, public feedback on these issues has become a crucial source of information. To analyze public sentiment in depth and provide valuable feedback to law enforcement, this study collected a total of 3,000 police-related video data through web scraping techniques. These videos encompass various scenes of police enforcement, policy interpretation, and social reactions, aiming to reveal the public's genuine attitudes and emotional responses toward police work through sentiment analysis. This insight is intended to help law enforcement better understand and respond to public opinions and needs.

During the dataset construction process, the collected videos were carefully screened and preprocessed to ensure data quality and representativeness. Specifically, the videos were categorized into five sentiment labels: Strong Positive, Weak Positive, Neutral, Weak Negative, and Strong Negative. This classification refines the depiction of public sentiment toward police enforcement and policies.After filtering and cleaning, 2,500 valid videos were obtained, with each category comprising 500 videos.

4.2 Parameters' Setting

In this study, to ensure the effectiveness and performance of the proposed model, experimental parameter settings covered multiple aspects, including data preprocessing, model training, and evaluation.

Data Preprocessing: Image data were resized uniformly to 224x224 pixels to meet the input requirements of the CLIP model. The pixel values were normalized to the [0, 1] range to ensure consistency of the input data. Text data were tokenized using the DeBERTa tokenizer, with the maximum sequence length set to 768 tokens to standardize the input format. Text features were extracted using DeBERTa's pre-trained word embeddings. Audio data were resampled to a 16kHz rate, and features were extracted using the Wav2Vec model.

Model Training: The learning rate was set to 0.0001, with a learning rate scheduling strategy employed to dynamically adjust the rate based on training progress. The batch size was set to 32 to balance training speed and memory usage. The AdamW optimizer was used to cater to the training needs of the model. The cross-entropy loss function was employed to handle the multi-class sentiment classification task. To prevent overfitting, regularization techniques such as Dropout layers and L2 regularization were applied.

Evaluation: Accuracy was used as the primary evaluation metric to quantify the model's performance in the fine-grained sentiment classification task.

4.3 Experimental Results

To validate the effectiveness of the proposed fine-grained sentiment analysis model for public opinion videos based on Conformer and multi-layered interaction attention (DW-MIACon) in multimodal feature fusion, a series of comparative experiments were conducted. These experiments systematically evaluated the performance of different fusion methods in multimodal sentiment analysis tasks.

In the first set of experiments, several representative multimodal interaction models were selected as baseline models for comparison with DW-MIACon. The specific models and their interaction mechanisms are described as follows:

TFN[17]: Tensor Fusion Network (TFN) uses a triple Cartesian product to decompose unimodal features into tensors and computes the outer product between modalities, capturing high-order interactions among features.

MFN[3]: Memory Fusion Network (MFN) leverages gated memory networks and attention mechanisms, using gated units to capture dynamic interactions among multimodal features over time, enhancing the model's ability to represent dependencies between modalities.

LMF[18]: Low-rank Multimodal Fusion (LMF) is an improvement over TFN, employing low-rank tensor decomposition techniques to enhance the efficiency and

performance of multimodal interactions while reducing computational complexity.

MulT[8]: Multimodal Transformer (MulT) utilizes directional cross-modal attention mechanisms to flexibly handle interactions between multimodal data at different time steps, implicitly addressing temporal alignment issues between data.

MMIM[19]: Multimodal Mutual Information Maximization (MMIM) introduces a hierarchical mutual information maximization framework that guides the model to learn shared representations across modalities, enhancing collaborative learning between them.

DW-MIACon: Dynamic Weighted Multi-layered Interaction Attention with Conformer (DW-MIACon) adopts a dynamic weighting strategy, adjusting the weight of each modality according to its importance at different stages of the task. Through multi-layered attention mechanisms, it performs stepwise interactions of features from different modalities, finely capturing features from each modality.

The comparison results between the DW-MIACon model and the aforementioned baseline models are shown in **Table 1.**

As shown in **Table 1.**, the DW-MIACon model significantly outperforms other baseline models in terms of accuracy, demonstrating its superior performance in multimodal feature interaction and fine-grained sentiment analysis. Specifically, the DW-MIACon model achieves an accuracy of 73.08%, which represents an improvement of 8.24, 6.02, 5.14, 4.36, and 2.26 percentage points over the TFN, MFN, LMF, MulT, and MMIM baseline models, respectively. These results highlight the significant advantages of DW-MIACon in capturing emotional features and interacting with multimodal data.

Table 1. Experimental Results of Sentiment Classification Model

Model		Accuracy
Baseline	TFN	0.6484
	MFN	0.6606
	LMF	0.6694
	MulT	0.6872
	MMIM	0.7082
Ours	DW-MIA	0.7308

Traditional interaction methods, such as TFN, perform feature interactions using a triple Cartesian product, capturing static relationships between modalities but lacking the capability to dynamically capture high-order features. MFN employs gated memory networks and attention mechanisms for modality interaction in the temporal dimension but struggles with complex, detailed emotional features across modalities. LMF improves computational efficiency through low-rank decomposition techniques, but its feature interaction richness is reduced, limiting its performance in fine-grained sentiment analysis. MulT enhances temporal alignment in multimodal interaction through directional cross-modal attention mechanisms but falls short in deeply integrating static features. MMIM enhances the shared and aligned representations of multimodal features by maximizing mutual information between modalities, yet it still lags behind DW-MIACon in capturing nuanced and deep emotional features.

The DW-MIACon model, with its dynamic weighting and multi-layered interaction attention mechanism, effectively enhances the fusion and emotional capture capabilities of multimodal features, allowing deep interaction at different levels. This mechanism not only facilitates information sharing between modalities but also dynamically adjusts the contribution of different modality features to sentiment classification, thereby improving the overall accuracy and robustness of the model.

In the second set of experiments, we compared the Conformer model with other commonly used multimodal fusion methods to further validate the effectiveness of Conformer in deep feature fusion. The comparison methods include simple concatenation (Concatenate), attention mechanism (Attention), and Transformer. The experimental results of various multimodal fusion models are presented in **Table 2.**

Table 2. Experimental Results of Various Models for Multimodal Fusion

Model	Accuracy
DW-MIA-Concatenate	0.6880
DW-MIA-Attention	0.6972
DW-MIA-Transformer	0.7128
DW-MIA-Conformer	0.7308

As shown in **Table 2.**, the DW-MIA-Conformer model significantly outperforms other commonly used fusion methods in terms of multimodal feature fusion accuracy, followed by Transformer, attention mechanism, and simple concatenation. Specifically, the DW-MIA-Conformer model achieved an accuracy of 73.08%, showing improvements of 1.80, 3.36, and 4.28 percentage points compared to DW-MIA-Transformer, DW-MIA-Attention, and DW-MIA-Concatenate, respectively.

The simple concatenation method straightforwardly combines features from different modalities; however, it fails to deeply model the interactions between modalities, limiting its fusion effectiveness and making it challenging to fully exploit the latent value of information from each modality. The attention mechanism improves the capture of key features but lacks in-depth modeling of complex inter-modal relationships, resulting in performance superior to Concatenate but still with substantial room for improvement. As a classic deep fusion approach, the Transformer effectively captures global relationships between modalities through self-attention mechanisms; however, its relatively weak ability to capture local dependencies impacts overall performance.

The Conformer model combines the strengths of convolutional and self-attention mechanisms, enabling it to capture both local dependencies and global feature information between modalities. This enhances the model's deep interaction capabilities, providing a better balance between local and global dependencies among modalities.

5. Conclusions

With the rapid development of social media and the widespread popularity of short video platforms, sentiment analysis faces increasingly complex emotional expression challenges. To deeply explore the intricate relationships between modalities and fully leverage the strengths of each modality in extracting comprehensive and profound emotional information, this study proposes a fine-grained sentiment analysis model for public opinion videos based on Conformer and Multi-layered Interaction Attention (DW-MIACon). Through refined and multimodal sentiment analysis, the DW-MIACon model can reveal emotions embedded within video content, providing more comprehensive and in-depth emotional insights for public opinion management.The DW-MIACon model enhances the performance of multimodal data in fine-grained sentiment analysis tasks through four main modules: feature extraction, modality interaction, modality fusion, and sentiment classification. Specifically, the feature extraction module employs DeBERTa, CLIP, and Wav2Vec models to extract high-quality features from text, image, and audio data, respectively. The modality interaction module utilizes a Dynamic Weighted Multi-layered Interaction Attention mechanism (DW-MIA), which performs multi-layered interactions and fusion of different modality features. This approach allows for dynamic weighting and adjustment of modality features, enhancing the collaborative effect between modalities. The modality fusion module employs the Conformer model for deep fusion, combining convolution and self-attention mechanisms to comprehensively capture both local and global dependencies between modalities, significantly improving the accuracy of sentiment classification.Experimental results demonstrate that the DW-MIACon model consistently outperforms existing methods in terms of accuracy, validating its effectiveness and advantages in multimodal fine-grained sentiment analysis.

Acknowledgments

This paper is supported by the Modeling and Prevention Mechanisms for National Strategic Network Public Opinion Risks in the Context of Major Changes (Project No. HB23ZT040).

References

[1] TANG X. Ethical Reflections on Media Emotional Reporting in the Post-Truth Era. Young Journalist, 2021, (12): 111-112.

[2] WILLIAMS J, KLEINEGESSE S, COMANESCU R, et al. Recognizing Emotions in Video Using Multimodal DNN Feature Fusion // Proceedings of Grand Challenge and Workshop on Human Multimodal Language:

Association for Computational Linguistics, 2018: 11-19.

[3] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory Fusion Network for Multi-View Sequential Learning. [2023-05-05]. https://arxiv.org/abs/1802.00927.

[4] ZADEH A, LIANG P P, PORIA S, et al. Multi-Attention Recurrent Network for Human Communication Comprehension // Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[5] HAN W, CHEN H, GELBUKH A, et al. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis// Proceedings of 2021 International Conference on Multimodal Interaction. New York, USA: ACM Press, 2021: 6-15.

[6] MORENCY L P, MIHALCEA R, DOSHI P. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web// Proceedings of the 13th International Conference on Multimodal Interfaces, 2011: 169-176.

[7] YU Y, LIN H, MENG J, et al. Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks. Algorithms, 2016, 9(2): 41.

[8] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal Transformer for Unaligned Multimodal Language Sequences// Proceedings of the Conference, Association for Computational Linguistics, 2019.

[9] YANG K, XU H, GAO K. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis// Proceedings of the 28th ACM International Conference on Multimedia, 2020: 521-528.

[10] FENG C, YANG H, WANG S, et al. Multimodal Sentiment Analysis Based on Top-Down Mask Generation and Stacked Transformers. Computer Engineering and Application, 1-11 [2024-09-10]. http://kns.cnki.net/kcms/detail/11.2127.TP.20240815.1136.002.html.

[11] WU J J, WANG J Y, ZHU P, et al. Dual-Modal Sentiment Computing Model Based on MLP and Multi-Head Self-Attention Feature Fusion. Computer Applications, 2024, 44(S1): 39-43.

[12] HE P, LIU X, GAO J, et al. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv preprint arXiv:2006.03654, 2020.

[13] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision// International Conference on Machine Learning. PMLR, 2021: 8748.

[14] BAEVSKI A, ZHOU Y, MOHAMED A, et al. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.

[15] CHEN Y, LAI Y B, XIAO A, et al. Multimodal Sentiment Analysis Model Based on CLIP and Cross-Attention. Journal of Zhengzhou University (Engineering Science), 2024, 45(02): 42-50. DOI: 10.13705/j.issn.1671-6833.2024.02.003.

[16] ZHOU J F, YE S R, WANG H. Text Sentiment Classification Based on Deep Convolutional Neural Network Model. Computer Engineering, 2019, 45(03): 300-308. DOI: 10.19678/j.issn.1000-3428.0050043.

[17] ZADEH A, CHEN M, PORIA S. Tensor Fusion Network for Multimodal Sentiment Analysis. arxiv preprint arxiv:1707.07250, 2017.

[18] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. arXiv preprint arXiv:1806.00064, 2018.

[19] HAN W, CHEN H, PORIA S. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. arXiv preprint arXiv:2109.00412, 2021.

