Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **4**, November **2025**

Analysis of the Complexity of College Students' Online Ideological and Public Opinion Based on Multimodal Deep Learning

Ming-Lan Sui*

Jilin Vocational College of Industry and Technology
Jilin 132013, P. R. China
15944296166@163.com

Chia-Huei Wu

Baekseok University
Seoul 06695, South Korea
fd6080@163.com

*Corresponding author: Ming-Lan Sui

Received September 5, 2024, revised April 21, 2025, accepted September 19, 2025.

Abstract. *With the rapid development of the Internet and the popularity of social media, college students' thoughts and public opinions have become more complex and diverse. Traditional public opinion analysis methods are difficult to fully capture and understand this complexity. This article constructs a multimodal complexity analysis model that combines time-domain convolutional networks with composite hierarchical fusion. The model adopts a composite hierarchical fusion mechanism. First, the three modal information of video, audio and text is dimensionally balanced, and mapped to the same dimensional space. Then, two-by-two fusion between different modes is performed to obtain three sets of dual-mode features. Finally, a feature matrix rich in multimodal information is obtained by fusing the three sets of dual-mode features. A residual operation is used to composite the original single-mode feature information and the three-mode information to obtain a multimodal feature matrix that is ultimately used for emotional orientation analysis. After each fusion, the time-domain convolution network is used to process its features, and finally, the soft attention mechanism is used to filter noise and redundant information, further improving the accuracy of the model. The experimental results show that the model is more accurate than traditional machine learning methods and provides valuable information for public opinion management.*

Keywords: multimodal; deep learning; sentiment analysis

1. **Introduction.** In the digital era, the development of the Internet and social media has greatly changed the way of information dissemination. As the most dynamic and creative group in society, college students are increasingly expressing their thoughts and opinions through online platforms. The diversity and immediacy of this information exchange make the ideological and public opinion of college students complex and varied, filled with rich emotions and perspectives. How to effectively capture, analyze, and understand this complex public opinion information has become an important issue that urgently needs to be addressed in both academic and practical fields [1, 2]. Traditional public opinion analysis methods, such as Support Vector Machine (SVM), Random Forest (RF), Naive Bayes Classifier (NBC), and Gaussian Mixture Model (GMM), mainly rely on single text data and are difficult to handle complex public opinion data containing

multimodal information such as video and audio [3, 4]. These methods have significant shortcomings in dealing with the diversity and complexity of data, making it difficult to comprehensively and accurately analyze the emotional tendencies and dissemination patterns of college students' online public opinion. To address this issue, this paper proposes a multimodal complexity analysis model based on time-domain convolutional networks combined with composite hierarchical fusion. This model innovatively adopts a composite hierarchical fusion mechanism. Firstly, the information of video, audio, and text modalities is dimensionally balanced and mapped to the same dimensional space. Then, pairwise fusion between different modalities is performed to obtain three sets of bimodal features. These bimodal features are further fused to generate a feature matrix rich in multimodal information. Finally, the residual operation is used to fuse the original single mode feature information and the three mode information, and the multi-mode feature matrix for emotional orientation analysis is obtained. After each fusion, the model uses a time- domain convolutional network for feature processing, and finally applies a soft attention mechanism to filter out noise and redundant information, thereby improving the accuracy of the model. The experimental results show that the model exhibits higher accuracy than traditional methods when dealing with complex multimodal data. This study not only provides more effective tools for public opinion management, but also offers new ideas and methods for the development of multimodal data analysis technology in the future. The following chapters will provide a detailed introduction to the model construction, experimental process, and result analysis of this study. Firstly, a detailed introduction will be given to the relevant theoretical techniques involved in this article, mainly focusing on the TCN and attention mechanism proposed in the article. Then, based on the theoretical foundation of deep learning methods, multimodal tasks were studied. Finally, the model was applied to analyze the complexity of college students' online ideological and public opinion, and the effectiveness of the proposed method was demonstrated through experiments.

1.1. **Related work.** Text is the modal feature that best expresses a person's mental state, as it can directly express emotional states through words. Early text analysis mainly relied on modern computer technology to extract subjective texts with emotional colors and analyze them. The current mainstream text sentiment analysis is based on deep learning technology for analysis and research. Speech sentiment analysis is a technology that studies human speech expression and its automatic recognition. In recent years, with the rapid development of speech technology and artificial intelligence technology, speech analysis has gradually become a popular research field. The current research status of speech analysis mainly includes the following aspects: (1) Dataset establishment: Research on speech analysis requires a large amount of speech data and emotional annotation. To address this issue, many researchers have started building their own speech datasets, such as IEMOCAP [5], Emo DB, RAVDESS [6], and so on. (2) Feature extraction: Feature extraction is an important step in speech analysis, aimed at extracting relevant features from speech signals. The commonly used feature extraction methods currently include MFCC, LPCC, Prosody, etc. (3) Recognition model: Recognition model is the core of speech analysis, and commonly used recognition models include SVM, decision tree, random forest, deep learning, etc. In recent years, research based on neural networks and deep learning algorithms has emerged in the field of speech analysis [7, 8]. The introduction and optimization of these algorithms not only improve the accuracy of speech recognition, but also make complexity analysis techniques perform better in practical applications. Image sentiment analysis refers to the recognition and analysis of emotions expressed in images through computer vision technology and machine learning algorithms.

The research in this field is mainly divided into two directions: one is emotion recognition, which identifies emotions expressed in images through image analysis; The second is emotion generation, which uses algorithms to generate images that conform to specific emotions. In recent years, research on image sentiment analysis has received widespread attention and in-depth exploration. The current mainstream image sentiment analysis methods include feature extraction based and deep learning based methods. Feature extraction based methods typically use manually designed features such as color, texture, shape, etc., which can reflect the emotional information contained in the image. And deep learning based methods usually use models such as Convolutional Neural Networks (CNN) to extract image features using their automatic learning ability, thereby achieving emotion recognition. In recent years, there have been many research advances in the field of image sentiment analysis both domestically and internationally. With the widespread application of deep learning algorithms, research in this field has gradually shifted from manual feature extraction methods to deep learning methods. Foreign researchers are leading in this area, mainly using convolutional neural networks for feature extraction and combining them with other deep learning algorithms for sentiment classification. For example, researchers at MIT in the United States [9] have implemented image sentiment classification using convolutional neural networks and recurrent neural networks (RNNs). Although significant progress has been made in some aspects of image sentiment analysis, it still has some limitations and challenges. Firstly, the interpretability of image sentiment analysis remains a challenge. Due to the black box nature of deep learning models, it is difficult to understand how the model produces prediction results. This uncertainty makes it difficult for researchers to determine how the model extracts emotional features from images and makes decisions, thereby reducing the credibility and reliability of the model [10]. Secondly, existing image sentiment datasets may have annotation errors and biases, which can affect the accuracy and generalization ability of the model. In addition, as emotions are subjective experiences, different people may have different emotional expressions towards the same image, which further increases the difficulty of data annotation. Thirdly, image sentiment analysis faces challenges across modalities and languages. Due to the fact that image sentiment analysis typically relies on text annotation and semantic understanding, there may be issues when dealing with cross linguistic and cross modal situations. In addition, image sentiment analysis also needs to consider the differences in culture and context in order to better understand and express emotions in images. Finally, due to the rapid development of computer vision and natural language processing, new technologies and algorithms continue to emerge, so image sentiment analysis needs to be constantly updated and improved to maintain its leading position in this field [11]. Multimodal sentiment analysis refers to the research field that utilizes multiple data sources (such as text, images, audio, video, etc.) to identify and analyze human emotional states. This sentiment analysis method can provide a more comprehensive understanding of human emotions, as different data sources can provide different information, thus complementing and enriching the results of sentiment analysis. Human emotional expression is complex and multidimensional. In communication, people not only use language and text, but also express emotional information through various ways such as changes in voice tone, facial expressions, and body movements. These media resources can serve as a basis for judging the current emotional state. Therefore, researchers gradually realized the limitations of analyzing emotional information through a single medium and began to integrate other media resources such as speech, audio, etc. into emotional analysis, developing research directions for multimodal emotional analysis [12]. Multimodal sentiment analysis not only improves the accuracy and precision of sentiment analysis, but also enables a more comprehensive understanding and recognition of the diversity and complexity of human emotional expression. Therefore, more and more researchers are paying attention to the study of multimodal sentiment analysis, which refers to the research field that uses multiple data sources (such as text, images, audio, video, etc.) to identify and analyze human emotional states. This sentiment analysis method can provide a more comprehensive understanding of human emotions, as different data sources can provide different information, thus complementing and enriching the results of sentiment analysis. Human emotional expression is complex and multidimensional. In communication, people not only use language and text, but also express emotional information through various means such as changes in voice tone, facial expressions, and body movements. These media resources can serve as a basis for judging the current emotional state. Therefore, researchers gradually realized the limitations of analyzing emotional information through a single medium and began to integrate other media resources such as speech, audio, etc. into emotional analysis, developing a research direction of multimodal emotional analysis. Multimodal sentiment analysis not only improves the accuracy and precision of sentiment analysis, but also enables a more comprehensive understanding and recognition of the diversity and complexity of human emotional expression [13]. Therefore, more and more researchers are paying attention to the research of multimodal sentiment analysis, such as image classification, object detection, natural language processing, etc. [14, 15]. In summary, with the continuous deepening of artificial intelligence research, multimodal sentiment analysis has also made tremendous progress. However, how to effectively utilize the interaction between single modal features and multimodal features for modeling is still the main problem faced by multimodal sentiment analysis. This article fully utilizes existing deep learning techniques to address the shortcomings of current multimodal sentiment analysis tasks, proposes corresponding solutions, and uses this model to analyze the complexity of college students' online ideological and public opinion, demonstrating the effectiveness of the proposed method.

1.2. **Contribution.** In response to the inability of traditional sentiment analysis methods to solve the complexity analysis problem of college students' online thoughts and public opinion, as well as the low accuracy and poor interaction between different modal information of existing multimodal sentiment analysis methods, a composite hierarchical fusion multimodal sentiment analysis model based on Temporal Convolutional Network (TCN) [16] was proposed through research on multimodal sentiment analysis methods. The main contributions are as follows: (1) In the process of multimodal emotional feature information fusion, the composite hierarchical fusion method is used. The individual modal information is fused in pairs to obtain three bimodal feature information, and the three bimodal information are fused with each other to obtain a feature matrix sequence containing three modal information. Finally, the similar residual connection method is used to fuse the above multimodal emotional information and three single-mode emotional information, so as to obtain a feature matrix with multimodal emotional feature information and strong interaction between different modes. (2) By utilizing the characteristics of dilated convolution and causal convolution in TCN networks, different modal feature information and the temporal features hidden behind the fused multimodal feature information are extracted. The obtained emotional feature information at different times is input into the TCN network layer for training, in order to avoid gradient vanishing or exploding caused by traditional RNN networks and solve context dependency problems. Add a soft attention mechanism at the end of the model. The purpose is to filter out redundant information noise and make the final analysis more accurate. (3) Multiple sets of experiments were conducted on the dataset to validate the fusion methods and overall model construction between different modalities mentioned above. Based on the experimental results, the model was optimized to evaluate the effectiveness of the multimodal feature fusion model and the accuracy of college students' online ideological and public opinion analysis.

2. Theoretical analysis.

2.1. Multimodal feature fusion.

Multimodal feature fusion refers to the process of integrating information from different modalities to obtain more complete and rich feature representations. The commonly used multimodal feature fusion methods include Feature Fusion, Decision Fusion, and Multi Channel Fusion. Feature fusion, also known as early fusion, refers to the fusion of multi-level features at an early stage, and then training a predictor on the fused features [17, 18]. Compared to traditional single-layer fusion methods, early fusion can better capture the interaction relationships between complex multi-level features, thereby improving prediction accuracy. Late fusion refers to first extracting and classifying individual features for each modality, and then fusing the classification results of each modality. This approach typically requires designing different models for each modality and aligning data between modalities for feature fusion and classifier training. Multi channel fusion: Multi channel fusion refers to the training of neural networks by inputting features from different modalities as different channels, in order to fully utilize the correlations between multimodal data. Multi channel fusion is more flexible compared to early fusion and late fusion, and can also effectively avoid the problem of excessive modal influence.

2.2. Time domain convolutional network.

TCN is a neural network structure based on CNN convolutional neural network, specifically designed for processing temporal data. Time domain convolutional networks are similar to traditional convolutional neural networks, consisting of convolutional layers, pooling layers, and activation functions [19]. But the convolution kernel size in time-domain convolutional networks is one-dimensional (i.e. convolving in the time dimension), and the same convolution kernel is used at each time step. This makes time-domain convolutional networks have good local correlation and translation invariance when processing temporal data. Each convolutional layer in a time-domain convolutional network can be seen as a certain degree of downsampling and upsampling of temporal data, in order to better learn long-range dependencies in time series. Moreover, in time-domain convolutional networks, information between different convolutional layers can be effectively transmitted and communicated, thereby further enhancing the network's ability to model temporal data. Compared to traditional recurrent neural networks and long short-term memory networks, time-domain convolutional networks have faster training speed and lower computational complexity [20]. In addition, time-domain convolutional networks can also accelerate the training and inference process through parallel computing. Time domain convolutional networks have been applied in various temporal data analysis tasks, including natural language processing, audio processing, and bioinformatics

(1) **TCN network structure.**

TCN mainly adopts two structures: causal convolution and dilated convolution.

Causal convolution: Causal convolution is a convolution operation in convolutional neural networks that takes into account the temporal causal relationships of time-series data. Unlike traditional convolution operations, causal convolution only allows convolution kernels to perform convolution from past to current time steps, and does not allow convolution kernels to perform convolution from current time steps to future time steps. The advantage of this design is that it can avoid the influence of future information on the current prediction results, ensuring the causality of the model. Therefore, causal convolution has certain advantages in processing time series data. Taking natural language processing as an example, if you want to predict the sentiment polarity of a word, only the preceding words can affect the sentiment of the current word, while the following words will not affect the current word. That is, it can only be calculated through the current input x_t and the previous input $x_1, x_2, \ldots, x_{t-1}$, which is a strict time constrained model and is therefore called causal convolution. The calculation formula is as follows:

$$P(x) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{1}$$

Dilated convolution: Dilated convolution, also known as atrous convolution, is a widely used operation in convolutional neural networks. It is a way to modify convolutional kernels by increasing their receptive field without increasing the number of parameters. Traditional convolution operations perform sliding window operations on the input tensor with a fixed step size, while dilated convolution adds some holes to the convolution kernel to increase the number of sampling points inside the kernel. These holes can be understood as adding some spacing on the input tensor, which is called dilation rate. When the dilation rate is 1, it is a traditional convolution. Dilated convolution involves inserting zero values in the middle of the convolution kernel to expand its receptive field and increase the number of sampling points in the input tensor. Dilated convolution can handle larger input tensors and preserve more original input information, making it suitable for tasks that require global information. For example, in image segmentation tasks, it is necessary to consider both global and local information simultaneously. In this case, dilated convolution can be used to extract a larger range of features. Unlike traditional convolution, dilated convolution allows for input skip sampling during convolution, i.e. using different sampling rates. The sampling rate of dilated convolution is controlled by parameter d, and the bottom sampling rate is 1, indicating that each point is sampled; The sampling rate of the middle layer is 2, which means that every 2 points, 1 point is sampled as input. As the number of layers increases, the size of the sampling rate also gradually increases. Therefore, the effective window size of dilated convolution increases exponentially with the number of layers, allowing the convolutional network to obtain a large receptive field with a small number of layers. In this way, convolutional networks can achieve efficient feature extraction with fewer layers.

(2) **Residual connection.**

Residual connections can alleviate the phenomenon of gradient vanishing or exploding to a certain extent, while TCN network structures can avoid this phenomenon through simple residual connections. The specific approach is to sum the input x and its non-linear mapped $G(x)$ to avoid the impact on the gradient caused by the increasing number of network layers. The dilation and causal convolution module adopted in this article normalizes the parameter hierarchy Hi-norm$(\cdot)$ after each dilation convolution calculation Conv$(\cdot)$, uses ReLU as the activation function for nonlinear calculation, and sums the results with the input to achieve residual parameter connection. The calculation formula is as follows:

$$R = x + G(x) \tag{2}$$

$$T_i = \text{Conv}(W_i \times F_j + b_i) \tag{3}$$

$$\{T_0, T_1, \ldots, T_n\} = \text{Hi-norm}(\{T_0, T_1, \ldots, T_n\}) \tag{4}$$

$$\{T_0, T_1, \ldots, T_n\} = \text{ReLU}(\{T_0, T_1, \ldots, T_n\}) \tag{5}$$

where T_i is the state value obtained by convolution calculation at time i; W_i is the matrix of words computed by convolution at time i; F_j is the convolution kernel of the j-th layer; b_i is the bias matrix; $\{T_0, T_1, \ldots, T_n\}$ is the encoding of the sequence after a complete convolution calculation. The TCN network layer expands the receptive field of convolution by stacking multiple dilated causal convolutional layers. A larger receptive field can obtain more complete sequence features, enabling the fused features to extract deeper semantic information. And enhance the information interaction between different modalities during the gradual fusion and extraction process, ultimately improving the overall performance of the model.

2.3. **Attention mechanism.**

The background of attention mechanism can be traced back to early research on face detection and recognition. In these tasks, people need to track faces in images and recognize them. To solve this problem, people need to find the most important face in the image and perform special processing on it. This problem can be transformed into how computers find attention centers in images. Attention mechanism is a core technology in machine learning, initially introduced in machine translation. In addition, it can also be used for tasks such as image classification and object detection, as it can find the most important objects in the image, thereby improving the accuracy of the model. In the field of computer vision, attention mechanisms have been widely applied [21]. For example, in image classification tasks, attention mechanisms can be used to enhance the visibility of the most important objects in the image, thereby improving the accuracy of the model. In addition, in object detection tasks, attention mechanisms can be used to find the most important targets in the image, thereby improving the accuracy of the model. In natural language processing tasks, attention mechanisms have also demonstrated their powerful effects. They can mimic the core idea that a certain word or phrase is the whole when people are reading or speaking, extract more useful information from large amounts of data, and automatically ignore unimportant information. In deep learning, attention mechanisms are mainly divided into two types: hard attention and soft attention. Hard attention refers to directly weighting a portion of input data to obtain an output result. This method usually requires manual parameter setting or training using reinforcement learning and other methods. Soft attention, on the other hand, calculates the weight distribution of all input data parts to obtain the output result. This method is usually implemented using the softmax function, which can automatically learn the weight of each input data. In text classification tasks, attention mechanisms can help models distinguish and classify different parts of text content. For example, for a sentence, attention mechanisms can be used to determine which words are more important for classification, thereby improving the accuracy of the model. Meanwhile, in machine translation tasks, attention mechanisms can help models model the relationship between source language and target language, improving translation quality. In short, attention mechanisms play an important role in deep learning. By focusing on key information, the performance and accuracy of the model can be improved, and it can also help the model process and classify complex input data more accurately [22].

3. **A Multimodal analysis model for the complexity of college students' network ideology and public opinion.**

3.1. **Composite hierarchical fusion.**

Dual modal fusion: For the fusion of different modal information, the single modal information is first fused pairwise. After pairwise fusion of the single modal information, three bimodal information are obtained, namely T+V (text+video), T+A (text+audio), and A+V (audio+video). The fusion process is shown in Figure 1.

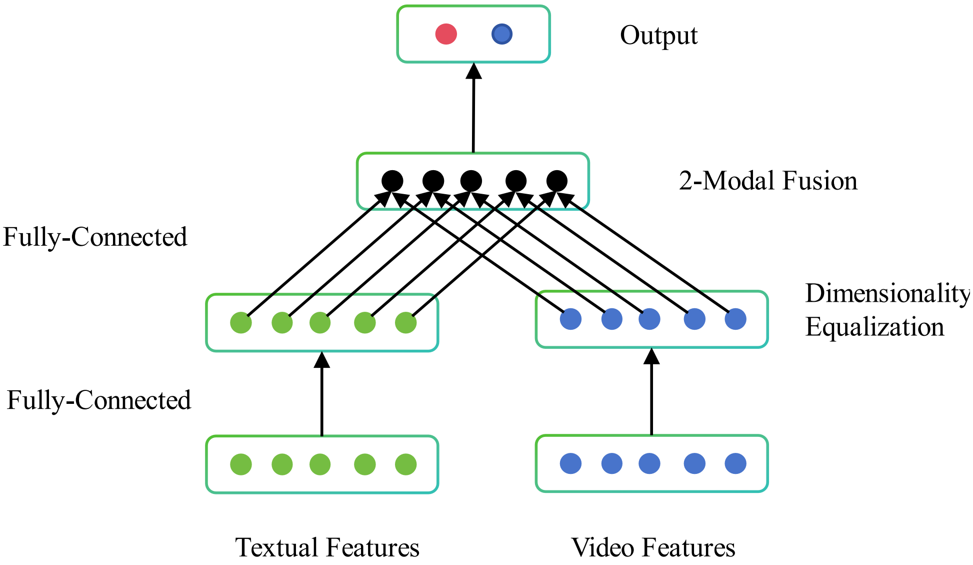

Figure 1. Diagram of bimodal information fusion

Three mode fusion: fuse the three bimodal eigenvectors obtained in the previous step to obtain a three mode eigenvector T+V+A, as shown in Figure 2.

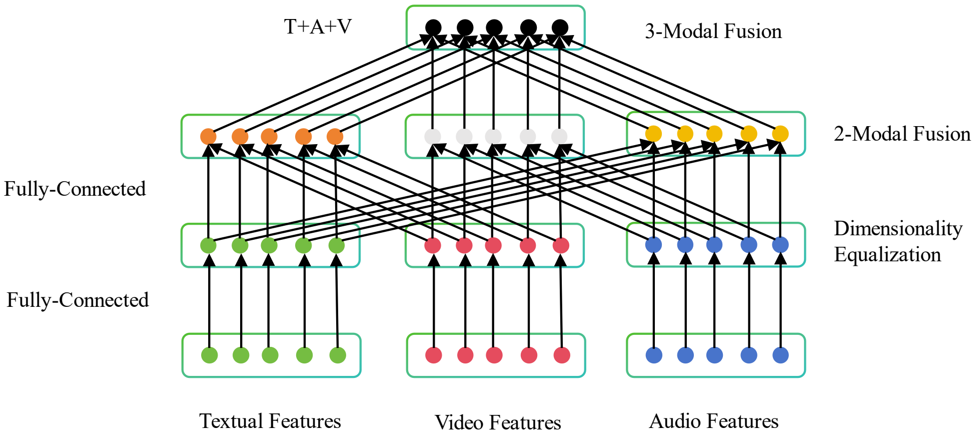

Figure 2. Diagram of trimodal information fusion

Composite fusion: on the basis of three modes of fusion, a structure similar to the residual network is used for composite level fusion, and its structure is shown in Figure 3. Experiments show that the fusion method using the composite level model can ultimately achieve better results in emotion classification.

3.2. **Multi modal feature information fusion process.** The overall structure of multimodal data fusion proposed in this article is shown in Figure 2, and the single modal sentiment feature vector is represented by the following entries, with the formula as follows:

$$\mathbf{f}_A \in \mathbb{R}^{N\times d_A} \tag{6}$$

$$\mathbf{f}_T \in \mathbb{R}^{N\times d_T} \tag{7}$$

$$\mathbf{f}_V \in \mathbb{R}^{N\times d_V} \tag{8}$$

The following terms are used to represent the single modal emotional feature vector. The formula is as follows: $\mathbf{f}_A$, $\mathbf{f}_T$ and $\mathbf{f}_V$ respectively represent the visual, textual, and

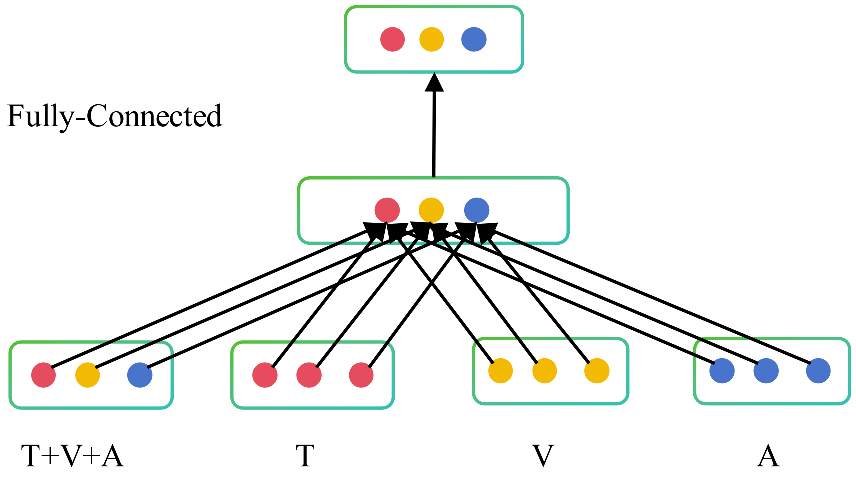

Figure 3. Diagram of hierarchical fusion

audio single modal feature information, and N is the maximum length of speech in the video. For shorter videos, use empty vectors of corresponding length to fill them with virtual discourse; For longer videos, perform corresponding cropping operations. In this article, $N = 50$. d_A, d_T and d_V respectively represent the feature dimensions of their corresponding modalities.

Single modal features $\mathbf{f}_A$, $\mathbf{f}_T$ and $\mathbf{f}_V$ have different dimensional features $d_A \neq d_T \neq d_V$, and before performing feature information fusion, they need to be mapped to the same dimension. In this model, it is mapped to $D_A = D_T = D_V = D$, D_A, D_T, D_V, which are the dimensions of the mapped single modal feature vectors such as video, text, audio, etc. After multiple experiments, it is found that the performance of the model is best when $D = 350$. The calculation process is shown in the following equation:

$$\mathbf{F}_A = \mathrm{ReLU}(\mathbf{f}_A W_A + b_A) \tag{9}$$

$$\mathbf{F}_T = \mathrm{ReLU}(\mathbf{f}_T W_T + b_T) \tag{10}$$

$$\mathbf{F}_V = \mathrm{ReLU}(\mathbf{f}_V W_V + b_V) \tag{11}$$

where $W \in \mathbb{R}^{d_A \times D}$, $b \in \mathbb{R}^D$, $W \in \mathbb{R}^{d_V \times D}$, $W \in \mathbb{R}^{d_T \times D}$, $b \in \mathbb{R}^D$, finally obtained $\mathbf{F}_A, \mathbf{F}_T, \mathbf{F}_V \in \mathbb{R}^{N \times D}$.

The three single modal feature information $\mathbf{F}_A, \mathbf{F}_T$ and $\mathbf{F}_V$ obtained above are used as inputs for the TCN network layer to extract single modal sequence features. The TCN network layer has been described in detail in Section 2.2 of this paper, and the calculation formula is as follows:

$$\mathbf{G}_A = \mathrm{TCN}(\mathbf{F}_A) \tag{12}$$

$$\mathbf{G}_T = \mathrm{TCN}(\mathbf{F}_T) \tag{13}$$

$$\mathbf{G}_V = \mathrm{TCN}(\mathbf{F}_V) \tag{14}$$

In order not to change the dimension of its output feature vector, the model also sets the number of output channels at the end of the TCN network layer to D. The dimension of the feature vector after passing through the TCN network layer is still $\mathbf{G}_A, \mathbf{G}_T, \mathbf{G}_V \in \mathbb{R}^{N \times D}$. From the above steps, the single modal feature information of audio, video, and text can be obtained. Next, the dual modal information fusion operation is performed, and the calculation formula is as follows:

$$\mathbf{f}_{VA} = \mathrm{ReLU}([\mathbf{G}_A \oplus \mathbf{G}_V] \cdot W_{VA} + b_{VA}) \tag{15}$$

$$\mathbf{f}_{VT} = \mathrm{ReLU}([\mathbf{G}_T \oplus \mathbf{G}_V] \cdot W_{VT} + b_{VT}) \tag{16}$$

$$\mathbf{f}_{TA} = \mathrm{ReLU}([\mathbf{G}_T \oplus \mathbf{G}_A] \cdot W_{TA} + b_{TA}) \tag{17}$$

where "·" represents matrix multiplication, "⊕" represents concatenation of two matrices, $W_{VA}, W_{VT}, W_{TA} \in \mathbb{R}^{2D\times D}$, and $b_{VA}, b_{VT}, b_{TA} \in \mathbb{R}^{N\times D}$. The model first concatenates two unimodal feature matrices, and then performs dimensionality reduction on the features. Finally, three bimodal feature matrices $\mathbf{f}_{VA}, \mathbf{f}_{VT}, \mathbf{f}_{TA} \in \mathbb{R}^{N\times D}$ can be obtained.

The obtained three bimodal feature matrices will be fed back into the TCN network layer as inputs for bimodal sequence feature extraction. The calculation formula is as follows:

$$\mathbf{F}_{VA} = \mathrm{TCN}(\mathbf{f}_{VA}) \tag{18}$$

$$\mathbf{F}_{VT} = \mathrm{TCN}(\mathbf{f}_{VT}) \tag{19}$$

$$\mathbf{F}_{TA} = \mathrm{TCN}(\mathbf{f}_{TA}) \tag{20}$$

Similarly, the dimension of the feature vector information after passing through the TCN network layer is still $\mathbf{F}_{VA}, \mathbf{F}_{VT}, \mathbf{F}_{TA} \in \mathbb{R}^{N\times D}$. Then use the same method to do three mode feature fusion. The fusion process is similar to that of dual mode fusion. The calculation formula is as follows:

$$\mathbf{f}_{VAT} = \mathrm{ReLU}([\mathbf{F}_{VA} \oplus \mathbf{F}_{VT} \oplus \mathbf{F}_{TA}] \cdot W_{TVA} + b_{TVA}) \tag{21}$$

where $W_{TVA} \in \mathbb{R}^{3D\times D}$ and $b_{TVA} \in \mathbb{R}^{N\times D}$, then take the fused three modal emotional features as the input of TCN network layer, and the calculation formula is as follows:

$$\mathbf{F}_{TAV} = \mathrm{TCN}(\mathbf{f}_{TAV}) \tag{22}$$

Finally, the composite level fusion is carried out, and the obtained three mode emotional feature $\mathbf{F}_{TAV}$ and single mode emotional feature $\mathbf{G}_A, \mathbf{G}_T$ and $\mathbf{G}_V$ are fused to obtain multimodal emotional feature vector. The calculation formula is as follows:

$$\mathbf{G}_{TAV} = \mathrm{ReLU}([\mathbf{F}_{TAV} \oplus \mathbf{G}_A \oplus \mathbf{G}_T \oplus \mathbf{G}_V] \cdot W_m + b_k) \tag{23}$$

In this model, a soft attention mechanism is used to input the obtained multimodal emotion feature vectors into the soft attention mechanism layer. The Softmax function is used to calculate the attention distribution matrix, and then the obtained attention distribution matrix is multiplied with the multimodal feature fusion matrix to obtain the final weighted multimodal feature matrix for the output of the final emotion classification result. The specific calculation process is as follows, and the calculation formula is as follows:

$$U = \tanh(\mathbf{G}_{TAV} W_1) \cdot W_2 \tag{24}$$
$$\mathrm{att} = \mathrm{softmax}(U) \tag{25}$$

$$\mathbf{F}_{\mathrm{scored}} = \mathbf{G}_{TAV} \otimes \mathrm{att} \tag{26}$$

where, $\mathbf{F}_{\text{scored}} \in \mathbb{R}^{N\times D}$ is the feature matrix obtained through the soft attention mechanism. $W_1 \in \mathbb{R}^{D\times D}$, $W_2 \in \mathbb{R}^{D\times 1}$ are weight matrices, "$\cdot$" represents matrix multiplication, and "$\otimes$" represents matrix element wise multiplication. This article ultimately uses multimodal emotional feature $\mathbf{F}_{\text{scored}}$ as the final for complexity analysis.

The model proposed in this article adopts a composite fusion method combined with TCN and soft attention mechanism. In the fusion process from single modal to dual modal and finally to the composite modal, multiple fusion extractions are performed, and after each fusion, the three modal information is tightly combined through the same TCN network, continuously improving the interactivity between different modal information in this process. All the obtained multimodal feature vectors are fed into the Soft attention mechanism for final filtering of redundancy and noise. During the processing of the attention mechanism, weak correlations can be weakened and strong correlations can be strengthened, thereby enhancing the interactivity between modal information.

4. Experiment.

4.1. **Experimental data.** During the period from January 1, 2023 to June 30, 2023, use web crawlers on the aforementioned social media platforms with keywords such as "university, doctoral student, deferred graduation, and paper publication". A total of 74330 samples were obtained, including 21556 microblog samples, 18583 WeChat samples, 15609 Tiktok samples, 7433 Zhihu samples and 11149 today's headlines samples. For the video text features (Text) in this article, transcription is first performed, and only Chinese transcription is used here. Add two unique markers to each transcript during transcription to indicate the beginning and end. Then, pre trained Chinese BERTbase word embeddings are used to obtain word vectors from the transcript. It is worth noting that due to the characteristics of BERT, this chapter did not use word segmentation tools. In the end, each word is represented as a 768-dimensional word vector dt=768. For the acoustic features (Audio) in the video, use the LibROSA speech toolkit to extract 22050Hz acoustic features with default parameters. Obtain 74 dimensional acoustic features da=74 in the MOSEI dataset. Extract visual features from the video at a frequency of 30Hz. Extract frames from the clip. This article uses the Multi Task Convolutional Neural Network (MTCNN) algorithm to extract aligned faces, and uses the MultiComp OpenFace2.0 toolkit to extract a collection of 68 facial landmarks, 17 facial action units, head posture, head direction, and eye gaze. Finally, 35 dimensional visual features with dv=35 were obtained in the MOSEI dataset. The ratio of dataset to test set to validation is set to 8:1:1, as shown in Table 1. Positive samples indicate public support or agreement, such as support for lifting strict restrictions on doctoral students publishing papers, appropriately relaxing journal requirements for publishing papers, and placing greater emphasis on the viewpoints presented in the paper rather than the authors. Negative samples indicate opposition or denial of petitioners' views on public opinion, such as questioning why petitioners should seek special treatment with lower standards because others can complete their work on time. These samples are used for model training.

Table 1. Dataset

Data set	Positive samples	Negative samples	Total samples
Training set	26623	32843	59466
Validation set	3327	4105	7432
Test set	3327	4105	7432

4.2. **Evaluating indicator.** To quantitatively describe the effectiveness of the model, accuracy and F1 score are used to evaluate the performance of sentiment classification. Precision refers to the ratio of the number of correctly classified samples in a category to the total number of samples in that category; The F1 value is the weighted average of precision and recall, and is a comprehensive indicator for judging the performance of a classifier. The accuracy and F1 value are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{27}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{28}$$

The confusion matrix is the foundation of various evaluation indicators, and the calculation of all evaluation indicators revolves around the confusion matrix. According to the confusion matrix, four indicator values can be obtained, among which TP is the true example, that is, the actual number of positive samples predicted as positive samples; FP is a false positive example, which refers to the number of negative samples that are actually predicted as positive samples.

4.3. **Comparison of experimental results.** In this section, P and F1 represent the accuracy and F1 score of the model in sentiment classification, respectively. The experimental comparison results of different models are shown in Table 2:

Table 2. Experimental comparison results of different models

Method	P	F1
EF-LSTM	77.84	78.34
TFN	78.50	78.96
LMF	80.54	80.94
MFN	78.94	79.55
MULT	81.15	81.56
MISA	82.67	82.12
TCN-CHF	84.12	84.46

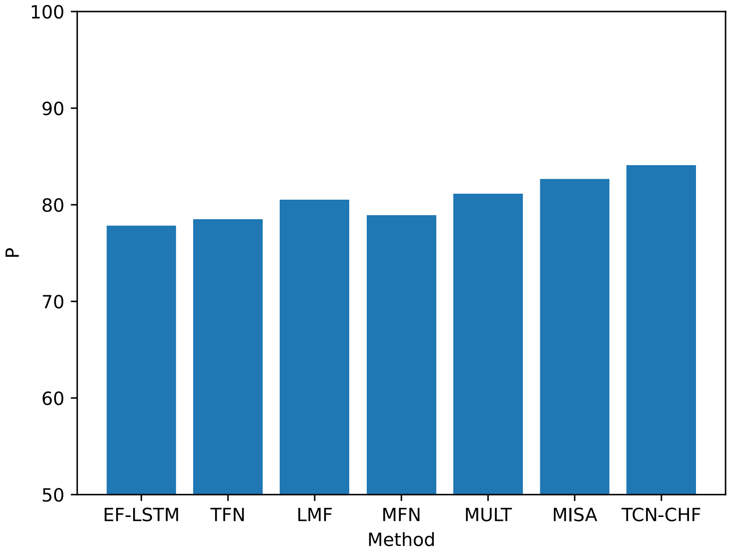

Figure 4. Comparison results of accuracy of different models

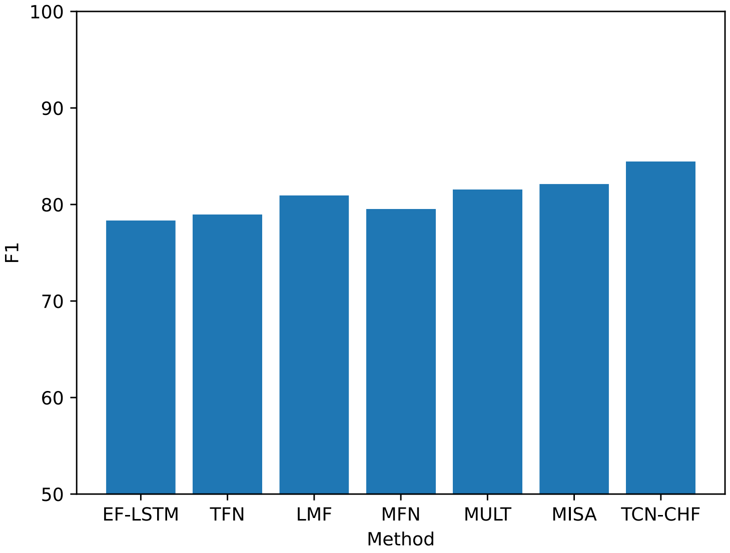

Figure 5. Comparison results of F1 score of different models

The experimental results in Figure 4 and Figure 5 demonstrate that the TCN-CHF model proposed in this paper performs better than other compared models, with an accuracy of 6.28 percentage points and an F1 score 6.12 percentage points higher than other models in sentiment classification. Especially compared to the advanced MISA model, the TCN-CHF model has improved accuracy by 1.45 percentage points and F1 score by 2.34 percentage points. This fully demonstrates the effectiveness and progressiveness of TCN-CHF model in multimodal emotion classification tasks.

5. **Conclusion.** This paper proposes a multimodal complexity analysis model combining time- domain convolutional network and composite hierarchical fusion mechanism, aiming to deal with the complexity and diversity of college students' ideological and public opinion under the Internet and social media environment. By balancing the dimensions of video, audio, and text modalities and fusing them pairwise, this model constructs a feature matrix rich in multimodal information for sentiment analysis. The use of residual operations and soft attention mechanisms further improves the model's ability to handle noise and redundant information. The experimental results show that this method outperforms traditional machine learning methods in accuracy and can provide more comprehensive and accurate support for public opinion management. This study has made innovative contributions to the field of analyzing complex online public opinion among college students and has important practical application value.

REFERENCES

[1] L. Chen, Y. Liu, Y. Chang, X. Wang, and X. Luo, "Public opinion analysis of novel coronavirus from online data," Journal of Safety Science and Resilience, vol. 1, no. 2, pp. 120-127, 2020.

[2] K. Richard, "Predicting the future with social media," The International Journal of Science in Society, vol. 3, no. 1, pp. 33-39, 2021.

[3] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," Computers, Materials & Continua, vol. 79, no. 1, pp. 19-46, 2024.

[4] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," Human-centric Computing and Information Sciences, vol. 9, 40, 2019.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, pp. 335-359, 2008.

[6] S. R. Livingstone, and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PloS One, vol. 13, no. 5, e0196391, 2018.

[7] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," Knowledge-Based Systems, vol. 161, pp. 124-133, 2018.

[8] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," Information Fusion, vol. 95, pp. 306-325, 2023.

[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, 2017.

[10] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," IEEE Transactions on Multimedia, vol. 25, pp. 3375-3385, 2022.

[11] 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," Knowledge-Based Systems, vol. 258, 110021, 2022.

[12] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "A survey of computational approaches and challenges in multimodal sentiment analysis," International Journal of Computational Science and Engineering, vol. 7, no. 1, pp. 876-883, 2019.

[13] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2276-2289, 2022.

[14] K. Kim, and S. Park, "AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis," Information Fusion, vol. 92, pp. 37-45, 2023.

[15] B. Yang, B. Shao, L. Wu, and X. Lin, "Multimodal sentiment analysis with unidirectional modality translation," Neurocomputing, vol. 467, pp. 130-137, 2022.

[16] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," IEEE Access, vol. 8, pp. 61672-61686, 2020.

[17] K. Zhang, Y. Geng, J. Zhao, J. Liu, and W. Li, "Sentiment analysis of social media via multimodal feature fusion," Symmetry, vol. 12, no. 12, pp. 2010, 2020.

[18] S. K. Panda, A. K. Jena, M. R. Panda, and S. Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," Multimedia Tools and Applications, vol. 82, no. 27, pp. 42763-42781, 2023.

[19] J.-J. Wang, C. Wang, J.-S. Fan, and Y. Mo, "A deep learning framework for constitutive modeling based on temporal convolutional network," Journal of Computational Physics, vol. 449, pp. 110784, 2022.

[20] Y. Wang, L. Deng, L. Zheng, and R. X. Gao, "Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics," Journal of Manufacturing Systems, vol. 60, pp. 512-526, 2021.

[21] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," Computational Visual Media, vol. 9, no. 4, pp. 733-752, 2023.

[22] G. Cheng, P. Lai, D. Gao, and J. Han, "Class attention network for image recognition," Science China Information Sciences, vol. 66, no. 3, pp. 132105, 2023.