Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-9, September 2025
DOI: 10.53469/jrse.2025.07(09).07

Research on Public Opinion Sentiment Analysis Based on Multi-modal Feature Fusion

Wenxin Fang, Tao Ye, Qinlong Xu

Qinghai Minzu University, Xining, Qinghai, China

Abstract: ***Under the background of rapid development of Internet technology, social media provides diversified expression channels for the public, and users are more inclined to use a combination of text and pictures to post comments. However, most of the current sentiment analysis methods use single-modal analysis, resulting in limited accuracy. To overcome this problem, this paper constructs a multimodal sentiment analysis model based on InceptionClip-Bert. First, text sentiment features are extracted with the help of the Bert model, and the Clip model is improved to extract image sentiment features; then, the cosine similarity is used to calculate the correlation between graphic and text sentiment tendencies to realize feature fusion; finally, public opinion analysis is carried out in terms of word frequency, word cloud, IP of the information publisher, identity of the information publisher, and T-SNE image clustering. The experimental comparison results show that this method can significantly improve the accuracy of sentiment recognition and provides a new idea for the research of multimodal feature fusion for sentiment analysis of public opinion.***

Keywords: Opinion analysis, Sentiment analysis, Graphic fusion, Multimodal modeling.

1. Introduction

1.1 Background and Social Value of the Study

With the high-speed evolution of information technology, social media platforms represented by microblogs, WeChat, forums, etc. are flourishing. These platforms rely on Internet technology to realize the acquisition and dissemination of data and information, and are characterized by a high degree of openness, social freedom and user interaction. According to the data of China Research Institute, the global social network platform market scale will reach 175.443 billion U.S. dollars in 2024, with a year-on-year growth of 16.93%; China's market scale will jump from tens of billions of dollars in 2013 to more than 200 billion dollars in 2024, with a compound annual growth rate of 35.96% [1]. Take TikTok as an example, its inbound purchase revenue in the first quarter of 2024 increased by 26% year-on-year, exceeding $2 billion, demonstrating strong market competitiveness [2]. In addition, social media platforms such as Facebook and Twitter have become the main channels for the American public to obtain news. These platforms have subverted the traditional information dissemination pattern by virtue of their large user groups and diverse information distribution methods. For example, during the 2020 U.S. election, the number of political discussions and news reports on social media platforms surged, becoming an important source of information for voters [3]. Social media has not only become the core carrier of information dissemination, but also deeply intervened in the fields of social governance, economic operation and public emotion expression.

The massive user comments on social platforms contain rich emotional information, and although these data exist in the virtual cyberspace, they can have an important impact on the real world through the spread of fermentation. In the economic aspect, when the stock market meltdown event breaks out, panic will spread rapidly through social media, causing a large number of stockholders to continuously sell their stocks, thus causing a new round of stock market crash [4]. At the level of social governance, the promotion of positive events and good practices through the publicity and reporting of social media can guide the public to establish correct emotional values and promote the construction of spiritual civilization [5]. The state and the domestic and international business community have been continuously paying attention to and promoting research in the field of emotion recognition. China has promulgated a policy "to make emotion one of the 60 major scientific research problems that need to be broken through" [6] to encourage researchers to innovate in the field of emotion recognition. In the corporate sector, authoritative reports state that the emotion detection and recognition market size is expected to be USD 57.25 billion by 2024 and is projected to reach USD 139.44 billion by 2029, growing at a CAGR of 19.49% during the forecast period (2024-2029) [7]. In foreign markets, in September 2022 CyberLink, a provider of artificial intelligence and facial recognition technology, integrated its FaceMe AI facial recognition engine into MediaTek's new Genio AIoT platform. Thanks to the recent integration of MediaTek's Genio 1200, FaceMe now has accurate facial recognition capabilities that can be flexibly deployed across different industries and use cases, including security, smart banking, access control, public safety, and smart retail [8]. Therefore, conducting sentiment analysis of social data has important research value and broad application prospects for social, economic, and political fields.

1.2 Status of Research and Problem Formulation

Currently, there is a large amount of useful information in online public opinion, as well as pornographic, violent, reactionary and other undesirable contents, whose uncontrolled proliferation may jeopardize the psychological health of individuals and threaten social stability and national security. In the case of the 2020 epidemic, for example, the influence of negative emotions in netizens' discussions is often stronger than that of positive emotions, and if it is not guided in a timely manner, it may lead to a crisis of public trust and affect social stability. In this context, it has become an urgent need for social governance to accurately analyze the emotional tendency of online public opinion.

Social media information presents multimodal features, and

users often express their emotions by combining graphics and text (e.g., using emoticons to reinforce emotions and pictures to convey irony). However, most existing studies rely on single-modal features, which lacks a comprehensive understanding of contextual information, and it is difficult to effectively capture potential inter-modal associations during multimodal fusion, resulting in limited model performance. Based on this, this paper proposes a multimodal sentiment analysis method based on early feature fusion to improve the accuracy of sentiment analysis.

2. Related Work

2.1 Text-Based Sentiment Analysis

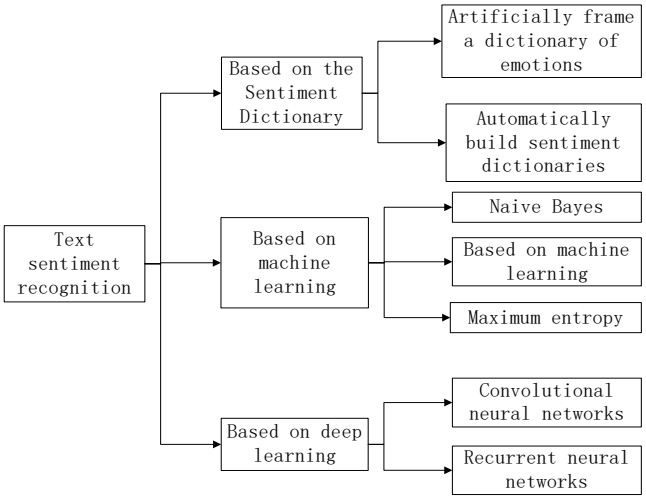

Figure 1: Text sentiment analysis method.

Text sentiment analysis methods mainly include three categories of methods based on sentiment dictionaries, machine learning and deep learning, as shown in Figure 1. Early studies were mostly based on manually constructed or automatically extended sentiment dictionaries, combined with rules or statistical methods to achieve classification. For example, Lee et al. [9] (2010) constructed an open-source sentiment corpus and proposed a recognition method based on a sentiment dictionary [10]; Wang et al. [11] (2013) expanded a sentiment dictionary in the field of education through the Laplace smoothing algorithm to realize news sentiment classification. Cui et al. [12] (2014) applied the theory of Support Vector Machines (SVM) to text categorization technology, and carried out a more in-depth research work on the issues related to the kernel function and parameters in SVM, which further improved the accuracy of the text categorization task. Jin et al. [13] (2015), on the basis of the SVM theoretical based on the lexical level, proposed a new feature representation called sentiment vector and applied the feature to the field of sentiment recognition using the traditional bag-of-words function, meanwhile, the authors also extracted the acoustic features corresponding to the text such as intensity, jitter, and spectral contour, which combined with late fusion finally achieved 69.2% recognition accuracy in the four classification experiments on the dataset they used.

With the rise of deep learning, neural network-based methods have become mainstream. kim [14] (2017) firstly applied convolutional neural network (CNN) to text feature extraction, capturing semantic features of text of different lengths through multi-scale convolutional kernels; xuejiao wang et al. [15] (2015) combined CNN and recurrent neural network (RNN), optimized the design of activation function, and improved the emotion recognition effect; Shelke et al. [16] (2022) proposed a deep network model, filtering irrelevant features through the feature ranking mechanism, and significantly improved the recognition accuracy. Although unimodal text analysis performs well in specific tasks, it lacks a comprehensive understanding of multimodal contexts.

To summarize, in unimodal detection, many studies still rely on single features of images or text for detection, and although they have achieved good results on some specific tasks, they lack a comprehensive understanding of contextual information.

2.2 Image-Based Sentiment Analysis

Classification tasks for image emotion recognition can be categorized into traditional methods and deep learning methods [17]. Since this study does not involve distributed learning tasks, only the classification task will be introduced. At present, classification tasks can be roughly categorized into traditional methods and deep learning methods according to the different ways of extracting image features.

Traditional methods rely on manual extraction of low-level visual features such as color, texture, composition, etc. [18] Machajdik et al. [19] (2010) identified image emotion through features such as color, texture, etc. Zhao et al. [20] extracted intermediate features based on the artistic principles of visual balance, harmony, and emphasis and used them for the prediction of image emotion. In addition, other researchers introduced adjective noun pairs (ANPs) into the field of ISR. Borth et al. [21] (2013) constructed the SentiBank visual concept detector to generate sentiment classification vectors based on adjective-noun pairs (ANPs). Chen et al. [22] generated sentiment classification vectors based on adjective-noun pairs (ANPs) through statistical analysis. six categories of objects with the highest frequency in the image and utilized the conceptual similarity between ANPs to build an emotion classification model. Rao et al. [23] used bags of visual words (bags of visual words) to extract features from each image patch to obtain emotion-related features. However, manual features are difficult to bridge the gap between low-level visual features and high-level emotional semantics.

With the breakthrough of CNN in several fields, more and more researchers tend to apply CNN to ISR. Based on the work of Borth et al, Chen et al. [24] (2015) proposed DeepSentiBank, which uses CNN to upgrade visual emotion concept classifiers. You et al. [25] utilized about 500,000 noisy images to train progressive convolutional neural network (PCNN) for emotion image classification. Rao et al. [26] utilized about 500,000 noisy images to train progressive convolutional neural network (PCNN) for emotion image classification. neural network (PCNN) for emotion image classification. Zhang et al. [27] Propose a multilayer image emotion recognition model, which consists of a bottom-level visual, a mid-level aesthetic, and a high-level semantic, and design a new loss function is designed to solve the problem of sample imbalance in emotion image dataset. The above work achieved some results but ignored the ability of local regions that can express emotions. Yang et al. [28] (2020) improved

the efficiency of local feature utilization by detecting emotional regions through the mechanism of region attention. Xiong et al. [29] proposed a region-based convolutional neural network using group sparse regularization for automatic detection of emotional regions. However, existing image analysis methods mostly ignore the semantic association between modalities.

2.3 Multimodal Feature Fusion

Considering the potentially great promise of multimodal emotion recognition work, more and more researchers have begun to synthesize common unimodal emotion information, such as audio, text, human physiological signals, visual information, and so on, to improve the accuracy of emotion prediction. Earlier studies such as De Silva and Chen et al. (2012) fused audio and visual features to demonstrate the effectiveness of multimodal fusion [30,31]. Meanwhile authors have also carried out more in-depth research work on fusing unimodal information such as audio and visual at both feature level fusion and decision level fusion, and although different fusion methods bring different gains they are all higher than the accuracy that can be achieved by using only any single input information. Eyben [32] (2013) used feature level fusion to fuse audio and textual information for emotion recognition work, again obtaining higher accuracy. Jing Chen (2016) combined various human physiological signals such as EEG signals, respiration, EMG signals and skin electrical responses for comprehensive analysis of human emotions and binary categorization of human emotions, with good results on a variety of publicly available datasets [33]. Poria et al. [34] (2017) applied attention for the first time within this area of research, by which the computation of attention to model the relationship between different modalities for fusion of different input information. Zadeh et al. [35] (2018) proposed the Multi-Attention Block (MAB) model to model text, audio, and video features simultaneously. Krishna et al. [36] (2019) further improved the performance of audio - text fusion through a cross-modal attention mechanism. Mittal et al. [37] (2020) applied a learning-based approach to recognize emotions across multiple input modalities and proposed a novel, data-driven multiplicative fusion method to combine different modalities by emphasizing more reliable cues and suppressing others on a per-sample basis, while using correlation analysis to differentiate between valid and invalid information, respectively, and by using proxy features for the Invalid modalities are substituted. The final model is not only robust to noise, but also achieves 82.7% recognition on IEMOCAP, which is a 5% improvement over previous work. However, as the number of modes increases, the complexity of the feature space rises, and how to effectively capture the nonlinear correlation between modes still needs to be improved.

In summary, during modal fusion, when the input modal information increases it will greatly increase the difficulty of feature fusion, and the potential correlation between the modal information cannot be well captured, which results in the degradation of model performance.

3. Model Construction

The InceptionClip-Bert multimodal sentiment analysis model proposed in this paper consists of five parts: data collection layer, data preprocessing, text feature extraction layer, image feature extraction layer, and multimodal feature fusion layer, as shown in Figure 2.

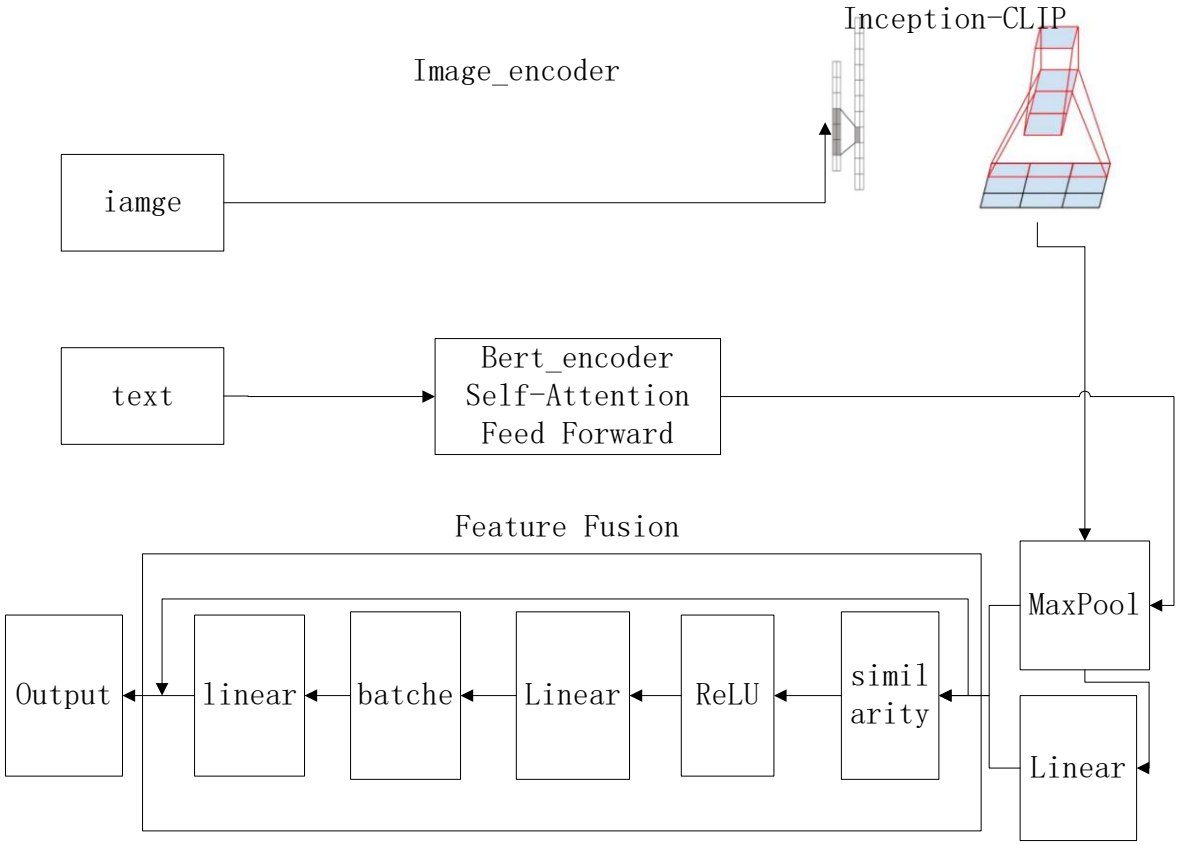

Figure 2: Architecture of InceptionClip-Bert multimodal sentiment analysis model.

3.1 Data Collection

Using Weibo as the data source, focusing on topics such as "artificial intelligence", "chat GPT", "AI", etc., requests crawler is used to collect data containing text, images, user attributes (e.g., IP, identity), and interaction data (e.g., number of retweets, number of comments, content of comments, time of comments, source of comments, number of likes of comments, id of the commenter, IP attribution of the commenter, name of the commenter, gender of the commenter, number of concerns of the commenter, number of followers of the commenter), and multimodal data. Specific steps include: setting keywords for advanced search, crawling microblog text, pictures, user information and comments (taking the first 60 pages of comments for each microblog), and constructing a dataset containing 17723 samples.

3.2 Data Pre-Processing

Text preprocessing: preprocessing includes operations such as initializing the predictor, data cleaning, emoji extraction, deactivation of words and Chinese word splitting, and visualization of data. First of all, data filtering: delete the data with empty comment content, delete the data with empty commenter IP attribution, delete the data with empty commenter gender, process the commenter IP attribution, remove emoji, print the cleaning results, and then use regular expression to extract all Chinese characters in the text, and then use the jieba thesaurus to split the Chinese words, and delete non-Chinese characters and deactivated words in the text. and delete non-Chinese characters and stop words in the text. Commenting on microblogs: the preprocessing process is mainly data cleaning, calling the "comment content" column in the data box of the deactivated word processing, and storing the processed results in the new column "deactivated comment content". Image preprocessing: Firstly, preprocessing is carried out, including resizing and normalizing the image.

3.3 Text Emotion Feature Extraction

The BERT model is used to extract the deep semantic features of the text. BERT solves the problem of polysemous word ambiguity by fusing word vectors, position vectors and sentence vectors through the bi-directional Transformer architecture. The word vectors of BERT are mainly composed of three vectors, namely, the word vectors, the position vectors of the words in the sentences, and the position vectors of the sentences in the individual text. In the case of polysemous words, this combination can effectively solve the problem of inaccurate model prediction. The Encoder model consists of three layers: query_layer (used to measure the relevance of the information in all other positions), key_layer (which determines which input positions are the most important for the current query), value_layer (the "content" associated with each position), which is used to measure the relevance of the information in each position to the current query. "), respectively, correspond to the proportional dot product attention module inside the multi-head attention mechanism. The internal Scaled Dot-Product Attention module (Scaled Dot-Product Attention) uses dot product for similarity calculation, and its detailed structure is shown in Figure 3.

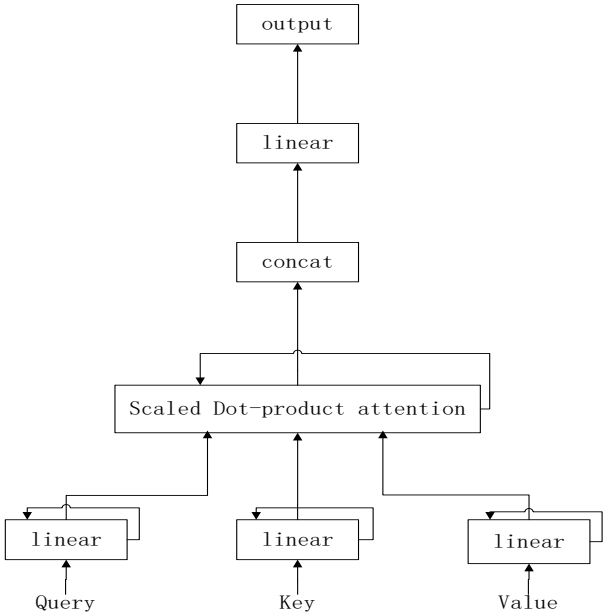

Figure 3: Multi-attention mechanism model.

Among them:

$$X \times W^{Q} = Q \quad (1)$$

$$X \times W^{E} = K \quad (2)$$

$$X \times W^{V} = V \quad (3)$$

Afterwards, by calculating the dot product of Q with all K, the calculated score values are used to measure the degree of attention of the rest of the input sentence to the word being encoded, and the result of the dot product is multiplied by a constant to limit the inner product to a controllable range; and then these scores are normalized to get the distribution of the attentional weights by the Softmax function, and the obtained results represent the magnitude of relevance of each word for the word in the current position The result represents the relevance of each word to the current position. Then the result obtained by Softmax is multiplied with V to get the value of Self-Attention at the current node (i.e., to get the context vector of each Query position):

$$\text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V = \text{attention}(Q, K, V) \quad (4)$$

Then the matrix obtained from BertLayer, is spliced, then normalized, and finally activated using the Tanh function.

The use of attention mechanism instead of the traditional Recurrent Neural Network (RNN) model is one of the highlights of the algorithm, which can effectively solve the problem of modeling the internal structure of sentences with the aid of pre-training on large-scale corpus by virtue of the multi-head attention mechanism.

3.4 Image Emotion Feature Extraction

The Inception_v3-CLIP network is proposed to replace the convolutional structure of CLIP with the Inception structure, and parameter optimization is performed to improve the efficiency of feature extraction.

We usually believe that the depth of the neural network plays a crucial role in the performance of the model, and as the

neural network deepens, the more nonlinear expressiveness, the more the model can learn. The key idea of the Inception model is to improve the accuracy by deepening the layers and depth of the network. The key to the Inception model is to stack convolutional kernels of different sizes together. which not only increases the receptive field but also improves the robustness of the neural network. Stacking different sized convolution kernels in a layer means that a layer can produce the effect of different sized convolution kernels, and it also means that instead of having to choose how to convolve the layer, the network learns what kind of convolution (or pooling) operation is best. Inception_v3 reduces the computational complexity by decomposing a 5×5 convolution into two 3×3 convolution operations, then decomposing all the convolution operations into two 3×3 convolution operations, and then decomposing all the convolution operations into two 3×3 convolution operations. Inception_v3 reduces the computational complexity by decomposing a 5×5 convolution into two 3×3 convolutions, and then decomposing all n*n convolution kernel sizes into 1×n and n×1 convolutions to speed up the computation.

The key idea of CLIP is to pre-train a neural network to learn a joint representation of images and their associated textual descriptions. It obtains supervised signals from text and uses contrast learning to create a pre-trained language-image model that is robust and scalable. To obtain perception from natural language, it trained a large model using more than 400 million pairs of data.

Specifically, the CLIP model consists of an image encoder f^1 and a text encoder f^T. CLIP classifies an image by computing the cosine similarity between the text cue p and the image features. Typically, a textual cue p is converted into a sentence, e.g., "a picture of a cat". Given an image x and the associated k category cues p_k, the prediction is computed by f^1 and f^T.

$$\hat{y}_{CLIP} = \text{argmax} < f^1(x), f^T(x) > \quad (5)$$

Where: k is the number of categories; <a,b> denotes cosine similarity, and strong cross-modal characterization is obtained by pre-training with 400 million pairs of data.

3.5 Feature

After obtaining the features of both text and image modalities, the problem of fusing the features of different modalities needs to be solved. The advantages and disadvantages of feature fusion will affect the accuracy of the final sentiment classification, and the common ways of fusing different modalities are splicing fusion, attention fusion, tensor fusion and so on. Splice fusion is to fuse different modal features of the same dimension through vector splicing technology, which is simple and convenient and saves computation time, but it will result in the loss of feature information, affecting the accuracy of sentiment classification; attention fusion is to introduce the attention mechanism before feature splicing, and to make up for the defects of the feature splicing fusion method by fully interacting with the information of different modal modes; tensor fusion is to fuse the different modal features through the tensor product, and the tensor fusion method will be more effective in the interaction of different modal features. Tensor fusion is to interact with different modal features through tensor product, and tensor, as a higher-order extension of vector or matrix, can fully explore the inter-modal features.

In this paper, we adopt a three-level fusion strategy of "graphic relevance calculation + attention weighting + tensor fusion":

1) Correlation calculation: the emotional tendency correlation between text feature X and image feature Y is measured by cosine similarity:

$$\text{similarity} = \text{cosine}(X, Y) = \frac{\sum_i^n (x_i \times y_i)}{\sqrt{\sum_i^n x_i^2} \times \sqrt{\sum_i^n y_i^2}} \quad (6)$$

2) Attention weighting: the mechanism dynamically adjusts the modal weights to generate weighted text features T_i and image features P_i:

$$T_n = W_T^T \cdot \alpha_i \quad (7)$$

$$P_n = W_P^T \cdot \alpha_i \quad (8)$$

Where, α_i denotes the attention weight value, generated by dot product operation.

The correlation, i.e., the attention weight, is calculated using dot product operation, and the attention weights are computed with the text features and image features separately to obtain the attention values of the two modal feature vectors.

3) attention weighting: fusion of cross-modal features by tensor product as shown in Eq:

$$M = P_n \otimes T_n \quad (9)$$

3.6 Opinion Analysis and Visualization

Text analysis: Generate word frequency bar charts and word cloud maps based on de-stopped text to analyze high-frequency emotional words; calculate regional emotional tendency through pnRatio method, and visualize regional emotional distribution using pyecharts maps; count the proportion of emotional tendency of users of different genders, and draw classification bar charts.

Image analysis: use CLIP to extract image embedded features, use K-Means clustering and T-SNE dimensionality reduction, and draw a 2D scatter plot to visualize the image sentiment clustering results.

4. Experiment and Analysis

4.1 Data set and Experimental Setup

In this paper, we use the original micro-texts, comments, and images on the microblogging social platform as a dataset for the experiments. Each micro-text and comment may contain a paragraph of text and an image, totaling 17,723 entries. In order to measure the effectiveness of the proposed system in this paper, the paper is tested on the MVSA [17] dataset, which consists of two independent datasets, MVSA-Single and MVSA-Multiple. MVSA-Single contains 5129 text and image pairs on the social media platform, and each pair has a pair of annotation tags; while MVSA-Multiple consists of 19600 pairs of text and images on the microblogging platform,

each with a pair of annotation tags; and MVSA-Multiple consists of 19600 pairs of text and images on the microblogging platform, and each pair of comments has a pair of annotation tags. Multiple consists of 19600 text and image pairs, each with three pairs of annotation tags. These annotation tags are independent of each other, and each of them labels the text as Positive, Negative or Neutral, so there are inconsistencies in the annotation tags of the MVSA-Multiple dataset. In this paper, we will follow the following rules to deal with inconsistent sentiment tags: if there are two or more tags with the same polarity, take that polarity, and if each of the three polarities is different, remove the set of text-image pairs. Similarly, the judgments of text markers and image markers are also opposite to each other, so there will be inconsistent emotion markers, and it is impossible to judge the emotion tendency of the whole group. In this paper, we will follow the principle of removing text-image pairs with completely opposite markers, i.e., the group with text labeled as positive and image labeled as negative, and vice versa, and the group with positive (negative) and a neutral marker, which is noted as positive (negative).

In order to improve the generalization ability of the model in this paper, the samples are split into training set and test set in the ratio of 4:1. The initialization of parameters is shown in Table 1.

Table 1: Parameter initialization.

serial number	parameters	parameter value
1	Batch Size	128
2	learning rate	5e-5
3	Epochs	3
4	optimizer	Adam

4.2 Evaluation Indicators

In order to test the performance of the InceptionClip-Bert multimodal graphic sentiment analysis model proposed in this paper, Precision, F1 value is used as the evaluation index. The calculation is shown in Equation (10) to Equation (11):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{10}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Preciosion+Recall}} \tag{11}$$

Where TP denotes that the sample was labeled as positive affect and was predicted to be positive; FP denotes that the sample was labeled as negative affect and was predicted to be positive; and FN denotes that the sample was labeled as positive affect and was predicted to be negative.

4.3 Experimental Results and Comparison

In order to validate the effectiveness of the model proposed in this paper on the task of sentiment analysis, we compare the Clip-Bert model with several classical sentiment analysis models.

1) Methods for image modality:

Resnet-50 [37]: Deep convolutional neural network model, Resnet-50 has 50 convolutional layers.

Methods for text modal.
BiLSTM [38]: Bidirectional LSTM model is one of the commonly used models in natural language processing tasks.

Bert [39]: large scale pre-trained language models for sentiment analysis tasks.

TGNN [40]: Using graph neural network models for sentiment analysis.

Methods for image-text bimodality.

CNN-Mulil [41]: Uses two separate convolutional neural networks to learn textual features and visual features and connects the learned textual features and visual features as to another convolutional neural network input.

Se-MLNN [42]: Several pre-trained image methods are applied to extract sufficient image features and ROBERTa-Base is used to construct text features for multimedia sentiment tasks.

The results of the experimental comparison are shown in Table 2.

Table 2: Experimental comparison results.

serial number	mould	accuracy	F1 value
1	Resnet-50	0.7467	0.7098
2	BiLSTM	0.8012	0.7790
3	Bert	0.8111	0.7624
4	TGNN	0.8034	0.7180
5	CNN-Mulil	0.7120	0.6536
6	Se-MLNN	0.8533	0.7580
7	InceptionClip-Bert	0.8859	0.8658

4.4 Opinion Analysis and Research

In this paper, we use the InceptionClip-Bert algorithm mentioned above to study the sentiment and AI", "chat GPT", and "AI" related topics, to study the propagation characteristics of public opinion information. In this paper, we crawl the original tweets, comments, and images on Weibo social media platform as a dataset for experiments and public opinion analysis, and each tweets and comments may contain a paragraph of text and an image, totaling 17,723 items.

Within the topic areas related to "Artificial Intelligence" "Chat GPT" "AI", for the textual data involved, this study ranks the words in descending order based on their frequency of occurrence, the The bar chart of the ten words with the highest word frequency is drawn, as shown in Figure 4. This chart can intuitively and clearly present the core content that the public pays most attention to and mentions most frequently in these hot topics, thus providing powerful data visualization support for the subsequent text analysis and topic exploration, helping to deeply explore the focus of public interest and discussion hotspots in the related fields, and further revealing the direction of public opinion and social concerns in this field.

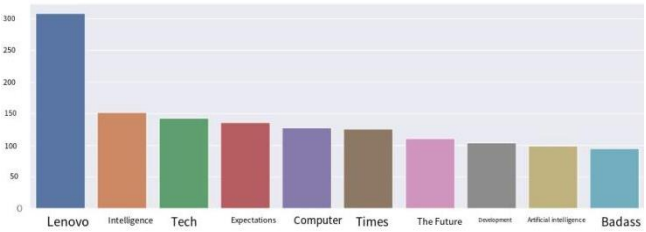

Figure 4: Bar chart of word frequency.

In this paper, we use pnRatio to quantitatively analyze the sentiment tendency in text data, and calculate the proportion of positive and negative sentiment in the discussion of "Artificial Intelligence" "Chat GPT" "AI" related topics in each region. We use pnRatio to quantitatively analyze the sentiment tendency in the text data, and calculate the proportion of positive and negative sentiments in the "Artificial Intelligence", "Chat GPT", and "AI" related topics in each region. With the help of pyecharts library, an intuitive map visualization model is constructed, as shown in Figure 5, which presents the proportion of emotional tendency in each region clearly and intuitively on the map, making the emotional dynamics of different regions clear at a glance. This visualization can accurately and timely capture the emotional tendencies expressed by netizens from various provinces on the microblogging platform in response to the above hot topics, providing valuable reference for the national and governmental departments in monitoring public opinion, helping them to take effective control measures at the early stage of the development of public opinion, guiding the general public to establish a rational and objective cognitive attitude, and creating a healthy and positive public opinion environment to better promote the development of "artificial intelligence" and the development of "artificial intelligence". This will better promote the smooth and orderly development and popularization of "artificial intelligence" and other emerging technologies in society.

Figure 5: Emotional tendencies by region.

Using the characteristics of the WordCloud library to generate word clouds, the word cloud generation process takes into account the word frequency, weight and other factors of the text, and the final generated word cloud is saved in the form of images, as shown in Figure 6, which presents the core points of the text and the distribution of key vocabulary in a unique visual way, and provides intuitive visual aids for the subsequent analysis of the text.

Further, this study used matplotlib, a powerful graphing tool, to carefully draw bar graphs to clearly visualize the percentage of different genders' comments on various types of emotional tendencies under the related topics of "Artificial Intelligence", "Chat GPT", "AI " and other related topics in terms of the percentage of comments on various emotional tendencies. Specifically, we clearly classify emotional tendencies into three categories: positive, negative and neutral, and for both male and female groups, we rigorously and meticulously counted the number of comments under each emotional tendency, and visualized them in bar charts, as shown in Figure 7. A visual presentation is carried out. The graph can make the difference in the distribution of comments of different genders in different emotional tendency dimensions clear at a glance, providing a powerful data visualization support for in-depth exploration of the role of gender factors in the public opinion field of the relevant topics, and helping to more accurately grasp the attitudinal tendencies of the different gender groups, thus providing a key reference basis for the subsequent research analysis and strategy development, and helping to achieve more targeted and effective This will provide a key reference basis for subsequent research analysis and strategy formulation, and

help realize more targeted and effective public opinion guidance and audience communication in related fields.

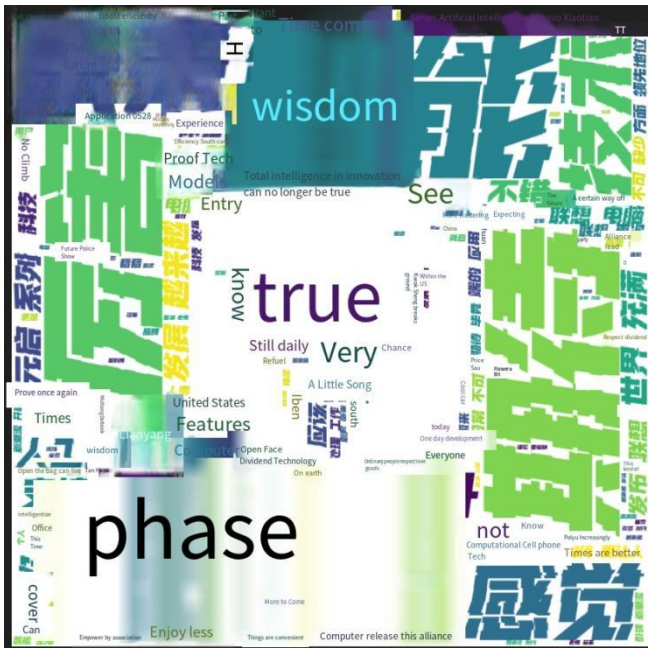

Figure 6: Word Cloud.

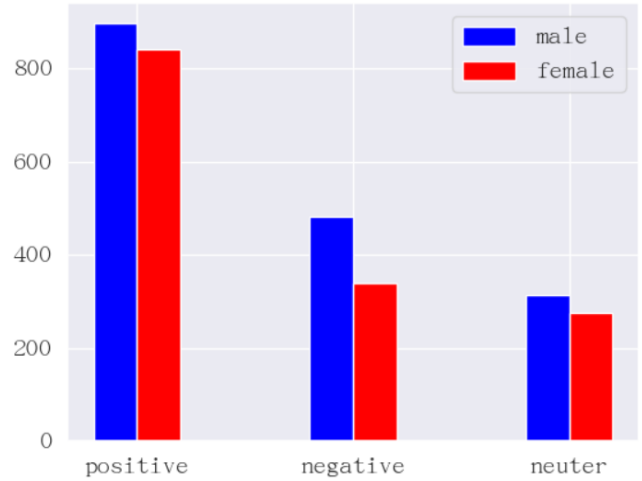

Figure 7: Tendency of men and women to comment on emotions.

In the processing of image data, the advanced CLIP model and its supporting processor are first introduced to realize the efficient loading of images, the accurate calculation of embedded features, and the effective execution of K-Means clustering analysis. Then, T-SNE, a cutting-edge algorithm, is used to subtly reduce the high-dimensional features obtained after clustering to a two-dimensional space, thus laying the foundation for subsequent visualization. Finally, the distribution characteristics of the data are visualized by drawing scatter plots, in which each sample point is distinguished by the corresponding color according to the cluster label it belongs to, and the center of the clusters is marked by an eye-catching red "X", and the specific effect is shown in Figure 8. The data processing and visualization process provides a clear, intuitive and reliable way of presenting the intrinsic structure and distribution law of the image data, which strongly supports the further development and in-depth exploration of related research work.

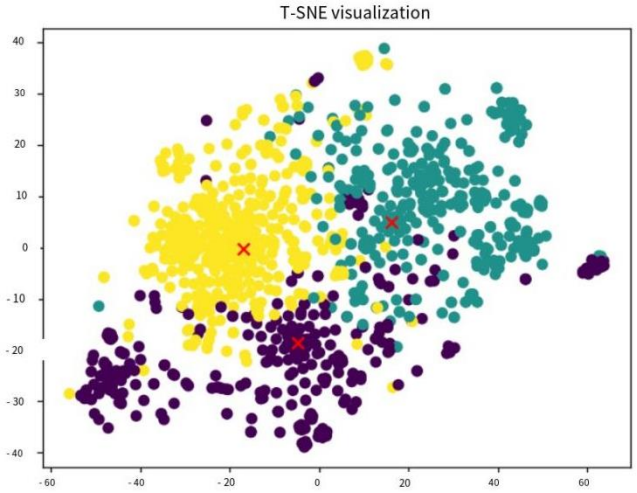

Figure 8: T-SNE visualization.

5. Concluding Remarks

In this paper, for the problem of multimodal sentiment analysis in social media, we propose a Clip-Bert-based graphic and text fusion model, which extracts the deep features of text through BERT, improves CLIP to extract the semantic features of images, and utilizes the attention mechanism and tensor fusion to realize modal interaction. Experimental results show that the model outperforms traditional methods in sentiment classification tasks, providing a new paradigm for public opinion analysis. Future research will further extend to multimodal data such as video and audio, optimize the cross-modal attention mechanism, and validate the generalization ability of the model in more fields.

Author Contributions: Conceptualization, Fang Wenxin and Ye Tao; methodology, Fang Wenxin and Wang Yuening; software, Fang Wenxin; validation, Fang Wenxin, Ye Tao and Wang Yuening; formal analysis, Fang Wenxin; investigation, Fang Wenxin; resources, Fang Wenxin; data curation, Fang Wenxin; writing—original draft preparation, Fang Wenxin; writing—review and editing, Fang Wenxin; visualization, Fang Wenxin; supervision, Fang Wenxin; project administration, Fang Wenxin; funding acquisition, Ye Tao. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The MVSA dataset is a public dataset available for download from the Multimedia Communications Research Laboratory.

Acknowledgments: This research could not have been completed without the assistance of many individuals, and I would like to express my deepest gratitude to all of them.

I would like to extend my sincere appreciation to Ye Tao, the co-author of this paper, for his professional insights in the research design. The viewpoints he proposed effectively optimized the research path. Wang Yuening also deserves my heartfelt thanks for his help in model construction, which has ensured the smooth progress of this research. I am also grateful to my laboratory colleagues. They selflessly shared

their experiences and offered valuable guidance when I encountered difficulties in experiments, and provided strong support during the data collection and collation stage. I would also like to thank other classmates for the inspiration brought by our daily exchanges and academic discussions. The strong academic atmosphere in the laboratory has always been the driving force for me to move forward.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SVM	Support Vector Machines
CNN	convolutional neural network
ANPs	adjective noun pairs
PCNN	progressive convolutional neural network

References

[1] Zhongyan Puhua Industry Research Institute. 2024-2029 China Network Computer Market Deep Research Report.

[2] Bellmann, L.; Hübler, O. (2021). Working From Home, Job Satisfaction and Work-life Balance-Robust or Heterogeneous Links. International journal of manpower, 42(3), 424-441.

[3] Beck, M.J.; Hensher, D.A.; Wei, E. Slowly coming out of COVID-19 restrictions in Australia: Implications for working from home and commuting trips by car and public transportation. *J. Transp. Geogr*. **2020**, *88*, 102846.

[4] Liang, Y.M.; Shen, Y.; Zhao, Y.Y. Emotional analysis of film reviews based on LSTM. *Digit. Commun. World* **2021**, *02*, 27-28.

[5] Yang, L.; Li, Y.; Wang, J.; Sherratt, R.S. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access* **2020**, *8*, 23522-23530.

[6] Liang, H.; Ganeshbabu, U.; Thorne, T. A dynamic Bayesian network approach for analysing topic-sentiment evolution. *IEEE Access* **2020**, *8*, 54164-54174.

[7] Lu, L.X.; Wu, D. Research framework of image emotion based on visual attention. *Doc. Inf. Knowl*. **2020**, *06*, 101-108.

[8] China Academy of Information and Communication Research. China Big Data Development Survey Report (2024).

[9] Feng, X. Facial expression recognition based on local binary patterns and coarse-to-fine classification. In Proceedings of the Fourth International Conference on Computer and Information Technology (pp. 178-183), Wuhan, China (16–16 September 2004).

[10] Lee, S.Y.M.; Chen, Y.; Huang, C.R. A text-driven rule-based system for emotion cause detection. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (pp. 45-53), Los Angeles, CA, USA (June 2010).

[11] Lee, S.Y.M.; Chen, Y.; Huang, C.R.; Li, S. Detecting emotion causes with a linguistic rule-based approach. *Comput. Intell*. **2013**, *29*, 390-416.

[12] Wang, Z.; Joo, V.; Tong, C.; Chan, D. Issues of social data analytics with a new method for sentiment analysis of social media data. In Proceedings of the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (pp. 899-904), Singapore (15–18 December 2014).

[13] Cui, J.M.; Liu, J.M.; Liao, Z.Z. Research on text categorization technique based on SVM algorithm. *Comput. Simul*. **2013**, *30*, 299-302.

[14] Jin, Q.; Li, C.; Chen, S.; Wu, H. Speech emotion recognition with acoustic and lexical features. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4749-4753), South Brisbane, QLD, Australia (19–24 April 2015).

[15] Chen, Y. Convolutional neural network for sentence classification. University of Waterloo, 2015.

[16] Li, Q.; Jin, Z.; Wang, C.; Zeng, D.D. Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems. *Knowl.-Based Syst*. **2016**, *107*, 289-300.

[17] Shelke, N.; Chaudhury, S.; Chakrabarti, S.; Bangare, S.L.; Yogapriya, G.; Pandey, P. An efficient way of text-based emotion analysis from social media using LRA-DNN. *Neurosci. Inform*. **2022**, *2*, 100048.

[18] Wang, Y.; Li, D.; Li, X.; Yang, M. PC-GAIN: Pseudo-label conditional generative adversarial imputation networks for incomplete data. *Neural Netw*. **2021**, *141*, 395-403.

[19] Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM International Conference on Multimedia (pp. 83-92), Firenze, Italy (October 2010).

[20] Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.S.; Sun, X. Exploring principles-of-art features for image emotion recognition. In Proceedings of the 22nd ACM International Conference on Multimedia (pp. 47-56), Orlando, FL, USA (November 2014.

[21] Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM International Conference on Multimedia (pp. 223-232), Barcelona, Spain (October 2013).

[22] Chen, T.; Yu, F.X.; Chen, J.; Cui, Y.; Chen, Y.Y.; Chang, S.F. Object-based visual sentiment concept analysis and application. In Proceedings of the 22nd ACM International Conference on Multimedia (pp. 367-376), Orlando, FL, USA (03 November 2014).

[23] Rao, T.; Xu, M.; Liu, H.; Wang, J.; Burnett, I. Multi-scale blocks based image emotion classification using multiple instance learning. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP) (pp. 634-638), Phoenix, AZ, USA (25–28 September 2016).

[24] Chen, T.; Borth, D.; Darrell, T.; Chang, S.F. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv:1410.8586* **2014**.

[25] You, Q.; Luo, J.; Jin, H.; Yang, J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1), Austin, TX, USA, February 2015).

[26] Rao, T.; Li, X.; Xu, M. Learning multi-level deep representations for image emotion classification. *Neural Process. Lett*. **2020**, *51*, 2043-2061.

[27] Zhang, H.; Xu, D.; Luo, G.; He, K. Learning multi-level representations for affective image recognition. *Neural Comput. Appl*. **2022**, *34*, 14107-14120.

[28] Yang, J.; She, D.; Sun, M.; Cheng, M.M.; Rosin, P.L.; Wang, L. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Trans. Multimed*. **2018**, *20*, 2513-2525.

[29] De Silva, L.C.; Miyasato, T.; Nakatsu, R. Facial emotion recognition using multi-modal information. In Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing (Vol. 1, pp. 397-401), Singapore (September 1997).

[30] Chen, L.S.; Huang, T.S.; Miyasato, T.; Nakatsu, R. Multimodal human emotion/expression recognition. In Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition (pp. 366-371), Nara, Japan (14–16 April 1998).

[31] Eyben, F.; Wöllmer, M.; Graves, A.; Schuller, B.; Douglas-Cowie, E.; Cowie, R. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *J. Multimodal User Interfaces* **2010**, *3*, 7-19.

[32] Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; Morency, L.P. Multi-level multiple attentions for contextual multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM) (pp. 1033-1038), New Orleans, LA, USA (18–21 November 2017).

[33] Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1), New Orleans, LA, USA (April 2018).

[34] Krishna, D.N.; Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In Interspeech (pp. 4243-4247), Shanghai, China (October 2020).

[35] Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 02, pp. 1359-1367), New York, NY, USA (April 2020).

[36] Niu, T.; Zhu, S.; Pang, L.; El Saddik, A. Sentiment analysis on multi-view social data. In MultiMedia Modeling: 22nd International Conference, MMM 2016 (pp. 15-27), Miami, FL, USA (4–6 January 2016).

[37] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778), Las Vegas, NV, USA (June 2016).

[38] Luo, Y.R.; Zhi, L.L. Word sense disambiguation in biomedical text based on Bi.STMIJ. *Softw. Guide* **2019**.

[39] Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal sentiment detection based on multi-channel graph neural networks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 328-339), Online (August 2021).

[40] Huang, L.; Ma, D.; Li, S.; Zhang, X.; Wang, H. Text level graph neural network for text classification. *arXiv:1910.02356* **2019**.

[41] Cai, G.; Xia, B. Convolutional neural networks for multimedia sentiment analysis. In Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015 (pp. 159-167), Nanchang, China (9–13 October 2015).

[42] Cheema, G.S.; Hakimov, S.; Müller-Budack, E.; Ewerth, R. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding (pp. 37-45), Online (27 August 2021).