International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7936-7948

https,//doi.org/10.70517/ijhsa463679

Machine Learning Based Sentiment Analysis of Social Network Users

Hao Zhang[1,*]
[1] Nanjing Forestry University, Nanjing, Jiangsu, 210037, China
Corresponding authors: (e-mail: zh18351267686@126.com).

Abstract The large amount of multimodal data generated by social network platforms contains rich sentiment information, and effective analysis of user sentiment is of great value for public opinion monitoring, business decision-making and user experience optimization. In this paper, we propose a cross-modal sentiment analysis model based on feature fusion, which extracts text sentiment features by BERT, image sentiment features by ResNet152, and adopts the multi-head attention mechanism to realize the effective fusion of graphical and textual information, and designs a feature-level fusion strategy to make full use of inter-modal correlation and independence information. The experiments are conducted on the Twitter public dataset, and the results show that the feature-level fusion method proposed in this paper improves the accuracy by 2.53% and 1.46% compared with the traditional feature splicing and Transformer fusion, respectively, and achieves 79.48% accuracy on the MVSA-Single dataset, which is 2.7% higher than that of the current popular DR-Transformer model by 2.7%. The simulation experiment selects the Paris Olympics event for multi-source social network user sentiment analysis, and obtains results that match the theoretical values in the three-source case of microblogging, WeChat and Sohu News, with a positive sentiment value of 8636.6 and a negative sentiment value of 2363.5, which verifies the validity of the model in the practical application scenarios. The study fully considers the mutual enhancement effect between graphic and textual modalities, solves the problem of insufficient accuracy of traditional single-modal sentiment analysis, and has important theoretical value and application prospects for social media sentiment monitoring, public opinion analysis, and personalized recommendation system.

Index Terms cross-modal sentiment analysis, multi-attention mechanism, feature fusion, social network, sentiment feature extraction, machine learning

I. Introduction

In today's online world, more and more people are willing to share their opinions and feelings on social networks. On social networking platforms such as Facebook, Amazon, and Yelp, review text data grows as the number of users and user-generated content grows, and online reviews can directly or indirectly influence future sales of e-commerce, restaurants, hotels, and other products [1]-[4]. Sentiment analysis for reviews can allow consumers to better judge the merits of a product, as well as allow product producers to make better decisions for future planning. For example, sentiment analysis of product reviews on online consumer platforms enables consumers to have a more comprehensive understanding of the information about the products they are interested in, so that they can formulate better plans for their own consumption plans, and enterprises can also understand market information based on the results of the analysis, so that they can optimize their products and sales strategies, so that user-generated text content has an important application value for many online and offline enterprises [5]-[8]. Sentiment analysis in comments, posts, videos, etc. published by social network users can detect and predict whether users have behaviors or tendencies such as depression and suicide, which can help to manage users' emotions and provide timely psychological intervention [9], [10]. In addition, for a certain social hotspot, by organizing and analyzing the related comments, it can allow functional departments to better formulate decision-making measures, such as monitoring public opinion on COVID-19, cyber violence, and child abuse, and digging out its patterns to ensure correct public opinion orientation, which can be used by the government as a strong basis for promulgating related policies [11]-[13]. Therefore, sentiment analysis plays an important role in social development and economic development. However, social networks constantly add a large amount of text data, which are not easy to be interpreted quickly. Finding information from social networks and analyzing it for sentiment by manual means has a very high degree of difficulty, and therefore smart technology-augmented sentiment analysis is needed.

According to the distinction of the methods of sentiment analysis, it can be divided into dictionary-based methods, machine learning methods and deep learning methods asked. Dictionary-based text sentiment analysis is based on

the sentiment dictionary, the provided sentiment words combined with linguistic knowledge to analyze the sentiment of the text [14]. Established machine learning methods mainly focus on training classifiers using a set of feature engineering methods such as Plain Bayes, Support Vector Machines [15], [16]. Early machine learning extraction of features relied on complex feature engineering and was a labor-intensive approach [17]. Recent studies have shown that deep learning provides new ideas for sentiment analysis [18]. Deep learning, named after its “deep” processing technique, uses backpropagation algorithms to learn features on successive model layers. Deep learning has superior performance for text feature extraction and has shown superior performance in the field of sentiment analysis [19], [20].

Nowadays, social media platforms have become an important channel for people to express their emotions and opinions, and the large amount of multimodal content generated every day carries rich emotional information. Effective analysis of such sentiment information can not only help enterprises understand users' attitudes toward products or services, but also provide government departments with an important basis for public opinion monitoring. Traditional sentiment analysis methods mainly focus on single-modal data, such as pure text or pure image analysis, but these methods often fail to comprehensively capture the sentiment features in cross-modal information. For example, a text with irony paired with a smiling image, analyzing any one modality alone may lead to biased sentiment judgments. Therefore, the development of sentiment analysis techniques that can effectively fuse multimodal information has become a hot issue in current research. In recent years, the rapid development of deep learning technology has provided new ideas to solve this problem. Pre-trained language models such as BERT perform well in text sentiment analysis tasks, while deep convolutional neural networks such as CNN and ResNet have achieved remarkable results in image feature extraction. However, it is still challenging to effectively fuse feature information from different modalities. Most of the existing feature fusion methods use simple splicing or weighted averaging strategies, which are difficult to fully exploit the mutual complementary and enhancement relationships between modalities. In addition, the importance of different modal information varies in different contexts, and more flexible fusion mechanisms need to be designed to adaptively adjust the contribution of each modality. As a typical multimodal task, social network user sentiment analysis has important practical application value as its accuracy directly affects the subsequent decision support and service optimization. In this study, a cross-modal sentiment analysis model based on feature fusion is proposed, which adopts BERT as a text feature extractor to capture the sentiment information in the text by using its powerful contextual semantic understanding; ResNet152 is chosen as an image feature extractor to extract the rich visual sentiment features in the image by deep residual network. In the feature fusion stage, the model introduces a multi-attention mechanism, which enables the features of different modalities to pay attention to and enhance each other, and designs a feature-level fusion strategy, which comprehensively utilizes the inter-modal correlation and the respective independent information. Experiments are conducted on a large-scale Twitter dataset to verify the effectiveness of the model and compare it with existing multiple graphic fusion sentiment classification models. In addition, the hot event of Paris Olympics is selected for simulation experiments to evaluate the performance of the model in the actual multi-source social network user sentiment analysis task, which provides technical support for the intelligent public opinion analysis system.

II. Relevant technology base

II. A. Visual Emotion Feature Extraction Techniques

The purpose of visual emotion feature extraction is to extract the visual information that can be associated with appropriate emotions. Different visual stimuli have a low to high effect on human emotions, such as color, texture, part, and object [21]. Methods for extracting these visual emotional features mainly contain traditional single visual feature extraction techniques and holistic feature extraction techniques based on deep learning.

II. A. 1) Traditional Visual Feature Extraction

(1) Gabor features and Wiccest features. Gabor filtering is sensitive to image edges and provides good orientation and scale selection properties. Wiccest features are mainly used to obtain color invariance within a region. Wiccest features are extracted by using natural image statistics to efficiently model texture information. The texture is described by the edge distribution of a particular image. Therefore, the histogram of the Gaussian derivative filter is used to represent the edge statistics. The complete range of image statistics in a natural texture can be well modeled using the complete Weibull distribution. This distribution is as follows:

$$f(r)=\frac{r}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma\left(\frac{1}{\gamma}\right)}\exp\left\{-\frac{1}{\gamma}\left|\frac{r-\mu}{\beta}\right|^{\gamma}\right\} \tag{1}$$

where r is the edge response of the Gaussian filter, $\Gamma()$ is the full gamma function, and $\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-1}dt$. β denotes the width of the distribution, γ denotes the peak height of the distribution, and μ denotes the mode of the distribution. The position of the mode is affected by uneven illumination and color illumination. Therefore, the value of μ is ignored in order to achieve color constancy.

The two-dimensional Gabor filter formed by the Gabor function has the property of achieving optimal localization in both the spatial and frequency domains simultaneously, and is therefore capable of describing the local structural information corresponding to spatial frequency, spatial location, and directional selectivity. The two-dimensional Gabon filter is as follows:

$$\tilde{G}(x,y)=G_{\sigma}(x,y)\exp\left\{2\pi i\begin{pmatrix}\Omega_{x0}\\ \Omega_{y0}\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}\right\} \tag{2}$$

where $G_{\sigma}(x,y)$ is the Gaussian distribution function with a standard deviation of σ, $\sqrt{\Omega_{x0}^2+\Omega_{y0}^2}$ denotes the radial center frequency, and $\tan^{-1}\left(\frac{\Omega_{y0}}{\Omega_{x0}}\right)$ denotes the direction.

(2) Color Emotion Feature Extraction

Color emotion can be described as the emotional feeling caused by a single color or a combination of colors, which is usually expressed by semantic words, such as warm, soft, active, etc. CIELab is a three-dimensional color space that represents the space arranged according to the three color appearance attributes, which contain hue, brightness and chroma. These three emotional dimensions are selected to establish an emotional color space, and the color map in the CIELab space is projected into the emotional space to get the emotional map, and factor extraction and coordinate transformation are performed according to the experimental results, and the coordinate transformation formula for each dimension is obtained as follows:

$$activity=-2-1+0.06\times\left(((L^*-50)^2+(a^*-3)^2+\left(\frac{b^*-17}{1.4}\right)^2\right)^{\frac{1}{2}} \tag{3}$$

$$weight=-1.8+0.04(100-L^*)+0.45cos(h-100°) \tag{4}$$

$$heat=-0.5+0.02(C^*)^{1.07}cos(h-50°) \tag{5}$$

where L^*,a^*,b^* are the CIELab coordinates, h is the CIELab hue angle, and C^* is the CIELab chromaticity.

(3) Global and Local RGB Histograms

Global RGB Histogram. Each cell of the histogram can be thought of as representing a single visual value, which in turn represents similar colors within a certain range. The size of the histogram bin indicates how many times a visual value appears in the image.

Local RGB Histogram. The global RGB histogram completely discards all information about the layout of colors in an image. If we first split the image into blocks and then compute the color histogram for each block, it is possible to obtain a preliminary descriptor of the colors at a rough location in the image, the local RGB histogram.

II. A. 2) Neural network feature extraction

With the advent of deep learning, we can now automatically extract visual features without prior intervention. Currently using CNNs to understand images is popular practice, the main advantage is that no relevant domain knowledge is required for visual emotion prediction, the specific approach is to first design and train the CNN network in the dataset, to get the pre-trained CNN network to extract the overall features of the image, which contain elements such as color, texture, objects, etc., the CNN network is generally composed of multiple layers with convolutional kernels.

II. B. Text-processing technology

II. B. 1) Word Embedding

The process of word embedding is to embed a high-dimensional space with the number of dimensions as the number of all words into a continuous vector space with a much lower number of dimensions. In this paper, we will introduce three main word embedding methods: word2vec, GloVe, and FastText.

(1) word2vec

The features of word2vec include: fast processing speed, semantically similar word vectors are also closer in similarity, and the word vectors obtained by word2vec satisfy semantic addition and subtraction [22]. word2vec has two main models, one is the Skip-gram model, and the other is the CBOW model. The Skip-gram structure is to use the intermediate words to predict neighboring words and CBOW model is to predict intermediate words using context words.

(2) GloVe

The core idea of GloVe, a method for training word vectors, is to obtain word representations by decomposing the 'word-word' co-occurrence matrix. Then an approximate relationship between the word vectors and the co-occurrence matrix is constructed and the representation of the word vectors is further learned using an objective function.

(3) FastText

FastText consists of three layers: input layer, hidden layer, and output layer. In the input layer, FastText uses character-level n-grams to represent each word, and employs a hash bucket approach to hash all the n-grams into the corresponding hash buckets, and all the n-grams that are hashed into the same bucket share a word vector. In the hidden layer, the input layer vectors are summed and averaged and multiplied by the hidden layer weight matrix, which is equivalent to summing the individual word vectors weighted as a vector for that sentence. The final output layer is obtained by multiplying the hidden by the output layer weight matrix.

II. B. 2) Text pre-processing

The text preprocessing method mainly includes the following steps:

(1) Noise removal

The collected text often contains unrecognizable, meaningless symbols, such as html tags, non-English characters and punctuation, etc., using regular expressions to remove such characters.

(2) Discontinued word filtering

Discontinued words are the words that are not necessary in a sentence, and removing them has no effect on understanding the semantics of the whole sentence.

(3) Stem extraction and pattern reduction

Stem extraction and word pattern reduction are exclusive to English texts. Both are used to find the original form of words. Only stem extraction is a little more aggressive, it may get not a word when extracting the stem.

II. C. Cross-modal fusion methods

In the field of cross-modal and multimodal sentiment analysis, the fusion of information between different modalities is the focus of research, which can often determine the effectiveness of cross-modal and multimodal sentiment analysis tasks. Currently, there are two main fusion methods: feature-level fusion and decision-level fusion.

(1) Feature-level fusion

Feature level fusion fuses the features extracted from visual features, text features and other modalities into a common feature vector, and combines these features for analysis. That is, the fusion of inter-modal information is carried out before the classifier, and its implementation based on deep learning generally has inter-modal feature splicing, addition, subtraction and multiplication.

(2) Decision level fusion

The decision hierarchy fusion method has to first detect and classify the features of each modality independently, i.e., there is an independent classifier for each modality, and the results are fused into a decision vector to obtain the classification results.

III. Cross-modal sentiment classification model based on feature fusion

This chapter proposes a multimodal emotion classification model based on feature fusion. The model architecture mainly consists of four parts: text emotion feature extraction, image emotion feature extraction, graphic emotion tendency correlation and graphic emotion feature fusion classification.

III. A. Text emotion feature extraction

Pre-training can be divided into two categories: context-independent and context-relevant. For context-independent models, such as Word2Vec, it will generate a word vector for each word in the glossary. In contrast, context-relative models will also generate a representation of each word based on other words in the sentence, such as the recently popular BERT model, which is widely used in context-relative work.

The word vectors of BERT are mainly composed of three vectors combined by summing, which are the word vectors, the vectors of the positions to which the words belong in the sentences and the vectors of the positions to which the sentences belong in the individual texts [23]. In the presence of polysemous words, this combination can effectively solve the problem of inaccurate model prediction. Its Encoder model mainly contains three linear layers: query_layer, key_layer, and value_layer, which correspond to the entry into the multi-head attention mechanism. Its internal proportional dot product attention module uses is the dot product for similarity calculation, and the model of the multi-head attention mechanism is shown in Figure 1. Among them:

$$X \times W^{Q} = Q \tag{6}$$

$$X \times W^{K} = K \tag{7}$$

$$X \times W^{V} = V \tag{8}$$

Afterwards, by dot-multiplying Q with K, the computed score value is used to measure how much attention the rest of the losing sentence pays to the word being encoded, the result of the dot-multiplication is multiplied by a constant to limit the inner product to a manageable range, and then computed by Softmax, and the result obtained represents the magnitude of the relevance of each word for the word at the current position.

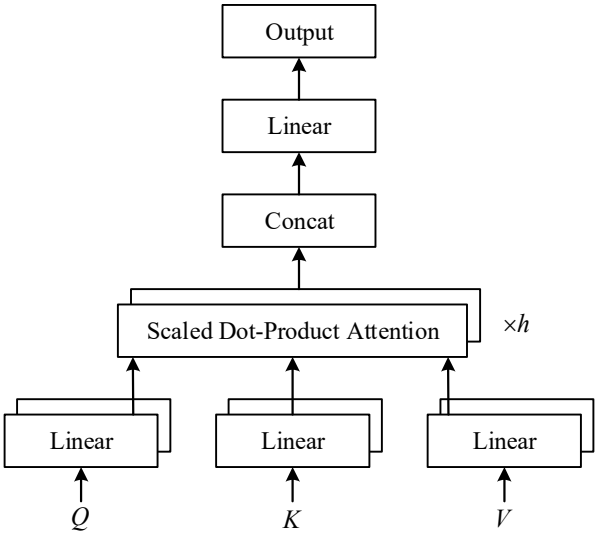


Figure 1: Multi-focus mechanism model

Then the result obtained from Softmax is multiplied with V to get the value of Self-Attention at the current node:

$$\text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V = Attention(Q,K,V) \tag{9}$$

Then the matrices obtained from the 12 sets of Bert Layer, are spliced, which facilitates the model to learn the relevant information in different subspaces, and then do the normalization process, and finally activate using the Tanh function.

The use of the attention mechanism instead of the traditional Recurrent Neural Network (RNN) model is one of the highlights of the algorithm. With the multi-head attention mechanism, aided by the pre-training of large-scale corpus, the algorithm can efficiently solve the problem of modeling the internal structure of the sentence.

III. B. Image Emotion Feature Extraction

In this paper, image features are extracted using ResNet152, which is a convolutional neural network (CNN) that utilizes a deep residual network (ResNet) to address the degradation of performance after deepening the network, where 152 represents the depth of the neural network [24].

The ResNet network is used as a reference for the VGG-19 network, which is modified to include residual units in the form of a short-circuiting mechanism, and the regular residual module is shown in Fig. 2. In practice, in order to save computation time and thus reduce the training time of the whole model, the bottleneck module will be used, and the dimension of the feature map will be cleverly scaled by using a 1×1 convolutional layer, so that the number of convolution kernels of the 3×3 convolutional layer is not affected by the previous layer, and its output will not affect the next layer, so that ultimately there is no effect on the accuracy of the model.

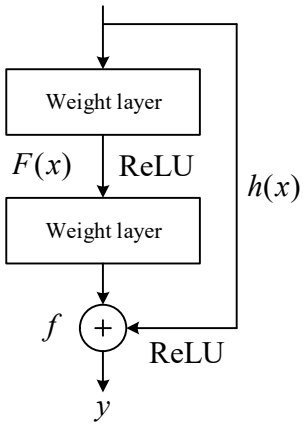


Figure 2: Schematic diagram of the residual module

III. C. Graphic Emotional Tendency Correlation

After extracting the textual emotional features and image emotional features, the textual features and image features are firstly spliced, and then the correlation of the graph asking emotional tendencies is calculated using cosine similarity:

$$similarity = cosine(X,Y) = \frac{\sum_{i}^{n}(x_i \times y_i)}{\sqrt{\sum_{i}^{n} x_i^2} \times \sqrt{\sum_{i}^{n} y_i^2}} \tag{10}$$

Cosine similarity is estimated by calculating the angle between two vectors to estimate the similarity.

Re-optimization splices text features, image features, text-image features, image-text features and uses Transformer coding for average pooling before calculating the correlation.

III. D. Text and Image Based Emotional Feature Fusion

The core idea of this paper is to merge textual emotion features and image emotion features in a more qualitative way, using the + structure instead of the traditional splicing structure, compared to the splicing structure that would merge the number of channels, i.e:

$$Z_{concat} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \tag{11}$$

where * denotes convolution, such a structure is mostly used in DenseNet, etc. However, + is the summation of the feature maps with a constant number of channels, i.e.:

$$Z_{+} = \sum_{i=1}^{c} (X_i + Y_i) * K_i = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_i \tag{12}$$

This approach undergoes feature variations so that the dimensions and features are all inside the same Euclidean space, and the amount of information under each dimension is increased, which saves more parameter and computation time, and is also obviously more favorable for the final classification.

In order to allow deeper networks to be trained well as well, this paper uses a residual structure for feature fusion, relying on the fact that the data output of one of the first number of layers is directly introduced into the input of a later data layer, so that the content of the later feature layer will have a linear contribution from one of the earlier layers. To analyze this problem from a mathematical point of view, the residual unit can be expressed as:

$$y_l = h(x_l) + F(x_l, W_l) \tag{13}$$

$$x_{l+1} = f(y_l) \tag{14}$$

where x_l and x_{l+1} denote the inputs and outputs of the l th residual unit, respectively, F is the residual function, $h(x_l)$ is a constant transform, and f denotes the ReLU activation function. It can be shown that the learning characteristics of the deep L are:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{15}$$

It is the residual learning structure of the short-circuited connected: that changes the computational nature of the affine transform from multiplicative to additive, from the point of view of backward propagation:

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_i) \right) \tag{16}$$

where loss denotes the cost function, so that it can be seen that $\frac{\partial loss}{\partial x_L}$ can be passed directly and the decay of the gradient is further suppressed.

IV. Experimental design and analysis of results

In this chapter, the selected Twitter public dataset is preprocessed, after which the cross-modal sentiment classification method based on feature fusion designed in this paper is validated based on this data, and the performance is compared with other models. Finally, the Paris Olympics event is selected for simulation testing to analyze the sentiment of social network users and verify the effectiveness of the sentiment classification model.

IV. A. Analysis of data sets

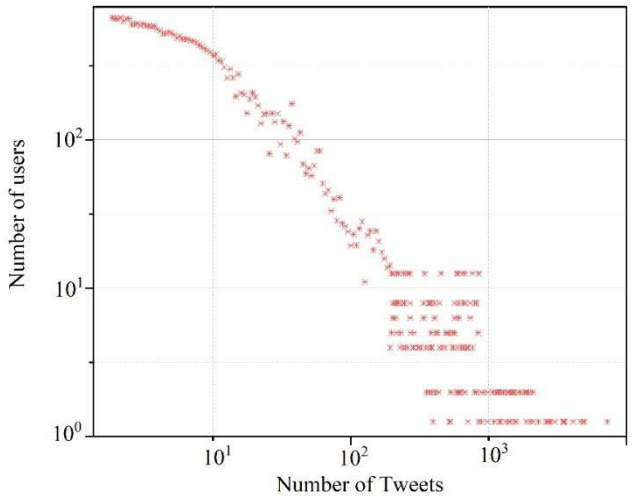
The original dataset used in this paper is derived from the SMP 2024 Twitter public dataset, which contains more than 30 million tweets data. After the text preprocessing operation, the statistical information of the experimental dataset obtained in this paper is shown in Table 1.

Table 1: Data set basic information

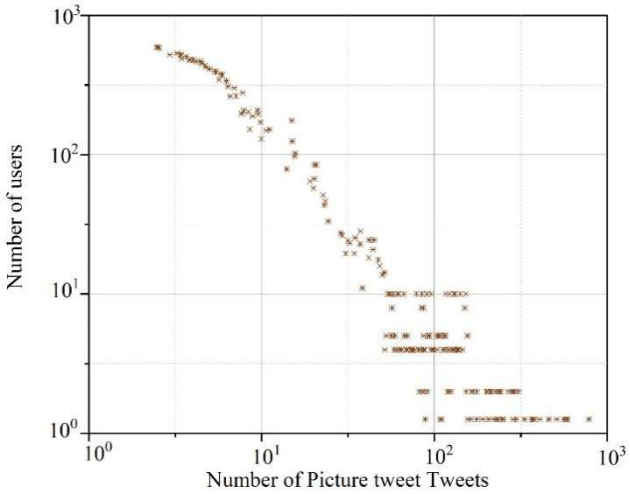
Statistical term	Quantity
Number of users	15346
Tweets	826431
Image tweets	116431
Comment number	254316
Text term	48613
Visual term	760

For this dataset, this paper first statistically analyzes its basic attributes, and the statistical analysis of the basic attributes of the dataset is shown in Fig. 3, where (a) to (c) show the frequency distribution of the total number of tweets posted by users in July 2024, the frequency distribution of the number of image tweets, and the frequency distribution of the number of comments obtained by tweets, respectively. The horizontal coordinate in the figure indicates the number of attributes corresponding to the object being counted, and the vertical coordinate indicates the frequency of the object being counted under a certain number. Taking (a) as an example, the horizontal

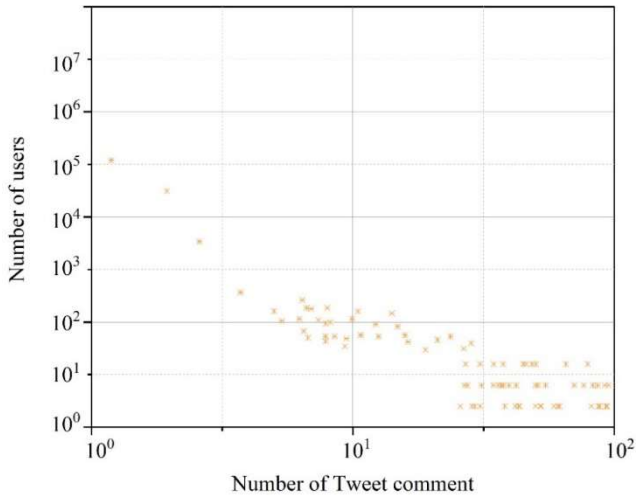
coordinate represents the total number of tweets posted by a user, and the vertical coordinate represents the number of users with that total number of tweets. From the figure, it can be seen that the basic attributes of the users in the dataset roughly conform to the power law distribution, which is consistent with the findings of related studies on existing complex social networks.



(a) Statistics distribution of total tweets



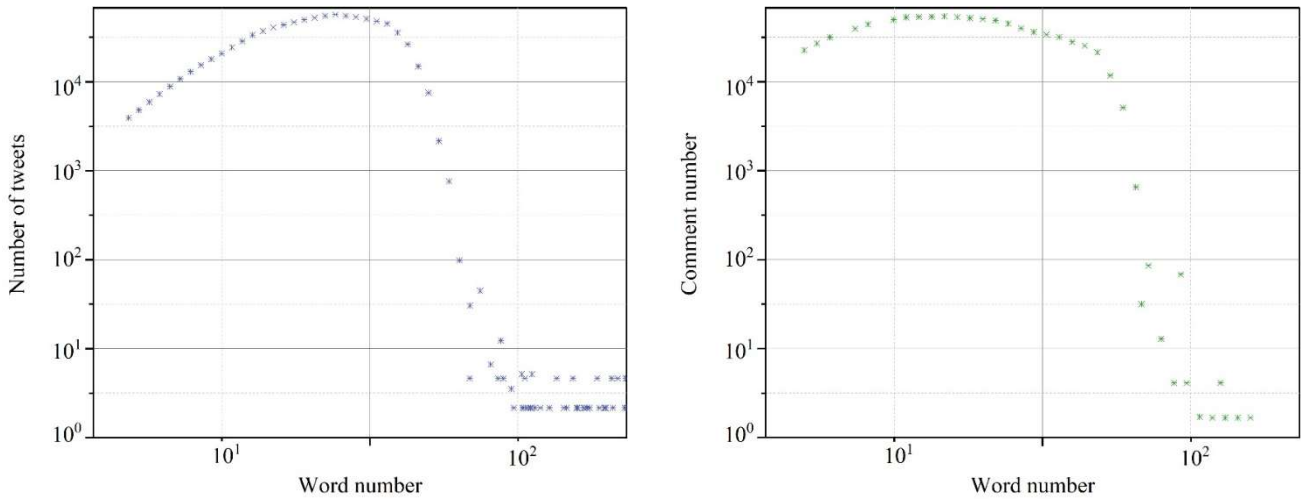
(b) The number of tweets in the user image



(c) Tweet review quantity statistics distribution

Figure 3: Statistical analysis of basic properties

For the tweets and comments in the dataset, this paper statistically analyzes their text lengths, and the statistical distribution of the number of text words for tweets and comments is shown in Fig. 4, (a) and (b) demonstrate the distribution of the number of text words for tweets and the distribution of the number of text words for comments, respectively. From (a) and (b), it can be seen that the text information when users post tweets and comments usually exists in the form of short text.



(a) Preprocessing the number of words in the previous article

(b) Review the number of words in the previous article

Figure 4: The statistical distribution of text words and comments

IV. B. Analysis of model evaluation

(1) Evaluation Metrics

In this paper, we use the following metrics to evaluate and validate our model: accuracy, recall, precision, and FI score. Their calculation equations are shown in equations (17) to (20). All evaluation metrics range from 0 to 100%, where the higher the value of the metric, the better the performance of the model.

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

$$recall = \frac{TP}{TP + FN} \tag{18}$$

$$precision = \frac{TP}{TP + FP} \tag{19}$$

$$F_1 = \frac{2 \times (precision \times recall)}{precision + recall} \tag{20}$$

where TP is the number of cases that are predicted to be positive and are indeed positive, TN is the number of cases that are predicted to be negative and are indeed negative, FN is the number of cases that are predicted to be negative but are actually positive, and FP is the number of cases that are predicted to be positive but are actually negative, where all other classes that do not belong to the current class are considered as negative cases.

(2) Evaluation results

In order to verify the effectiveness of the cross-modal sentiment classification method based on feature fusion, the following nine models are selected for comparison and validation on Twitter dataset. Among them, the textual feature extraction part uses the textual sentiment classification model RBCF. the VGG-16, ResNet-50 and ResNet152 models are used to extract features from image data. In the data fusion part, the direct splicing and Transformer, which are commonly used in existing research methods, are chosen to compare with the multi-head attention mechanism used in this paper. The final experimental results are shown in Table 2. From the experimental results of (3), (4) and (5), it can be obtained that the image feature vectors extracted by VGG-16 and ResNet-50 contribute less to the accuracy of the graphic fusion emotion classification task, and even will be lower than the accuracy of the image in a single modality. From the comparison of experimental results in (1), (2), and (8), it can be seen that the accuracy of the graphic fusion sentiment classification algorithm based on the multi-attention mechanism reaches 88.24%. Compared with the emotion recognition algorithm based on single modality of text and image data, it improves 9.92% and 6.58% respectively. This indicates that data of different modalities can play a complementary role to each other. Image features and text features can play an enhancing role to each other.

Therefore, the fusion features of image-text modality can achieve better classification results. From the comparison of experimental results (6), (7) and (8), it can be concluded that compared with the (6) method of direct splicing of feature vectors of different modalities, or the (7) method of sending sequences into the Transformer network for fusion after splicing, this paper can effectively capture the correlation between the modal features of the image and text based on the multi-head attention mechanism, and can carry out attention weighting of the features of different modalities to achieve adaptive fusion of features. The accuracy is improved by 2.53% and 1.46%, respectively. And the Macro-F1 value of the fusion method used in this paper is better than the other two methods, and the best results are achieved. Experiment (9) is the experimental results of feature-level fusion obtained by splicing the text output, image output, fused feature output, and the three results. Comparing from experiments (1), (2), (8), and (9), it can be seen that the accuracy of feature-level fusion is improved by 11.24%, 7.9%, and 1.32% compared to single text decision, image decision, and fused feature decision, respectively. This shows that feature-level fusion can well utilize the correlation information and independence information between modalities. And under the Macro-F1 evaluation index, feature-level fusion is significantly higher than other models, and the predictive ability of the model is well generalized for different categories of samples.

Table 2: The results of different emotional classification models were compared (%)

N	Model	Precision	Recall	F_1	Accuracy
(1)	RBCF	71.59	65.37	68.36	78.32
(2)	IRCF	76.72	74.68	75.69	81.66
(3)	RBCF+VGG-16	73.95	67.04	70.33	81.16
(4)	RBCF+ResNet-50	76.86	74.98	75.91	82.38
(5)	RBCF+ResNet152	77	75.2	76.09	84.81
(6)	RBCF+IRCF(CONCAT)	78.84	77.31	78.06	85.71
(7)	RBCF+IRCF(Transformer)	81.04	83.01	82.02	86.78
(8)	RBCF+IRCF(Multi-focus mechanism)	83.64	83.02	83.32	88.24
(9)	Ours	83.87	85.23	84.55	89.56

Considering the usage requirements and scope of the system, this paper focuses on model training and evaluation on graphic datasets containing Chinese text in the above research work. In order to further validate the effectiveness and robustness of this paper's approach and enhance the experimental persuasiveness, in addition to the experiments on the Twitter dataset, comparative experiments with other graphic-text fusion sentiment categorization models are also conducted on some other publicly available datasets.

When conducting experimental comparisons, in order to be able to make direct comparisons with cutting-edge technologies, it is necessary to ensure the fairness and reliability of the comparison results by conducting experiments and evaluations on the same datasets. Therefore, in this paper, two publicly available English graphic datasets collected by MCRLab Lab on Twitter, which are more widely used in cutting-edge technologies, are selected: the MVSA-Single dataset and the MVSA-Multi dataset.

The data is preprocessed and the word vectors are fed into the subsequent layers of the network for training, which can result in a text feature extraction model suitable for English. The proposed method in this paper is compared experimentally with the following models:

(1) CNN-Multi: This model uses two CNN networks for text and image feature extraction respectively, and directly splices image features and text features for sentiment classification.

(2) DNN-LR: This model uses DNN networks to classify text and images for sentiment classification, followed by a decision fusion approach using an averaging strategy.

(3) MultiSentiNet: uses VGG to extract target and scene information from images. The word vector representation is extracted using Glove, and the word vectors are input into the LSTM network to get the text feature representation.

(4) CoMN: The cross attention approach is used to obtain the key features of the image labeled using text and the key features of the text labeled using image, respectively.

(5) VistaNet: the model obtains text feature vectors by means of bidirectional GRU and attention mechanism and image feature vectors by means of VGG-16. The fusion of text and image is realized with the help of attention mechanism.

(6) MVAN: This model enhances the scene features, object features and text features with each other using the iterative attention mechanism, and the enhanced fused features are used for emotion classification.

(7) DR-Transformer: This model uses RoBERT and DenseNet to extract text and image information respectively, and adds the identification layer after directly splicing the two vectors to identify different modalities in order to differentiate the origin of the modalities, and then fuses the features through the Transformer Encoder.

The model comparison results are shown in Table 3. From the analysis of the experimental results, it can be seen that models (1) and (2) have the worst performance among all the methods, which is because models (1) and (2) do not take advantage of the correlation between different modalities, although they take into account the information of two modalities. Models (3) ~ (8) all consider the interactions between the graphic modalities. Among them, model (3) has the worst performance, which is due to the fact that the method only considers the unidirectional relationship of image to text. Model (4) considers the bidirectional relationship between image and text, which is an improvement compared to model (3). Model (5) solves the problem of modal alignment by utilizing Attention mechanism for feature fusion between image and text. Model (6) establishes the relationship between image and text from multiple perspectives. Model (7) has a greater improvement in the experimental results, which is due to the use of better performance unimodal feature extraction algorithm for model (7).

Table 3: Compare the results with his model (%)

N	Model	MVSA-Single		MVSA-Multi	
		Accuracy	F_1	Accuracy	F_1
(1)	CNN-Multi	61.19	58.37	66.29	64.2
(2)	DNN-LR	61.43	61.02	67.85	66.32
(3)	MultiSentiNet	69.84	69.62	68.86	68.1
(4)	CoMN	70.52	70.03	68.93	68.84
(5)	VistaNet	72.4	71.89	70.83	69.14
(6)	MVAN	72.98	73	72.36	72.29
(7)	DR-Transformer	76.78	73.5	75.1	73.21
(8)	Ours	79.48	75.64	77.79	74.33

In this paper, not only the unimodal feature extraction algorithm is optimized. In the modal fusion part, the correlation between different modalities and modality-unique information are considered. From the experimental results, it can be seen that the cross-modal emotion classification model based on feature fusion designed in this paper achieves superior performance compared with other models.

IV. C. Analysis of actual simulation results

The study selected the Paris Olympics event for simulation, and captured 100 articles from Sina Weibo, WeChat public website, and Sohu News for sentiment analysis respectively.

On the premise of fixing 1000 pieces of data from Weibo, the positive sentiment value is 2063.36 and the negative sentiment value is 723.6, varying the amount of data from the other two sources. The two source sentiment analysis image is shown in Figure 5. Sentiment analysis algorithm based on multi-source data can be obtained with a positive sentiment value of 2063.36 and a negative sentiment value of 723.6.

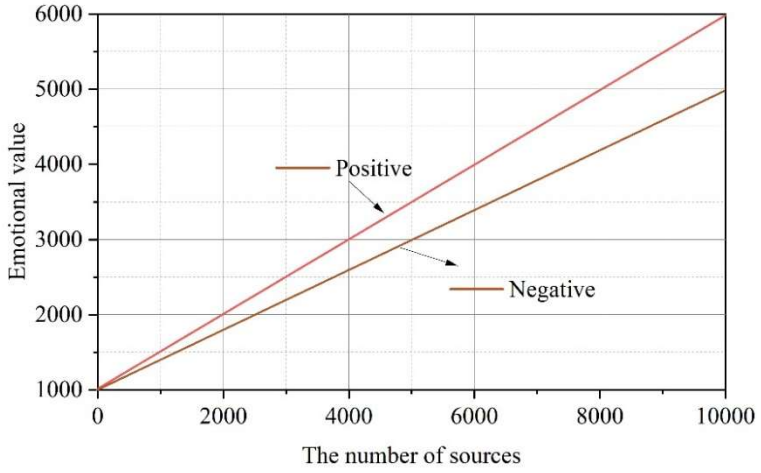
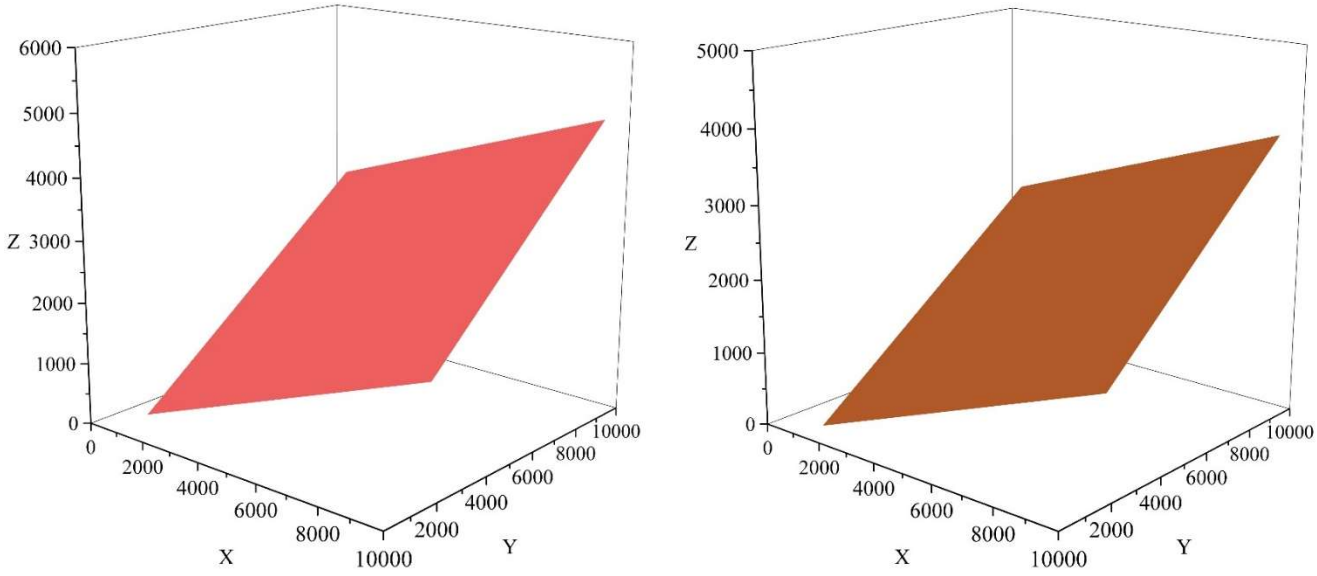


Figure 5: The two source emotion analysis image

The comprehensive sentiment value in the case of microblogging, WeChat and Sohu News is 8636.6 in the positive direction, and 2363.5 in the negative direction.The three-source sentiment analysis image is shown in Fig. 6, and (a) and (b) indicate the positive and negative sentiment analysis, respectively. Comprehensive image analysis of the theoretical value of the positive 8636.1, negative 2363.9, the algorithm results and the actual results of the basic charm match.



(a) Positive (b) Negative

Figure 6: Three source emotion analysis image

Based on the above analysis results, it can be shown that for multi-source social network public opinion, the cross-modal sentiment classification model based on feature fusion can effectively excavate the implied public opinion behind popular events and improve the novelty and accuracy of the public opinion analysis results in multi-source social network scenarios.

V. Conclusion

The effectiveness of multimodal fusion in social network user sentiment analysis has been confirmed through research and experiments on a cross-modal sentiment analysis model based on feature fusion. The experimental data show that the accuracy of the graphic and text fusion sentiment classification algorithm based on the multi-attention mechanism reaches 88.24%, which is 9.92% and 6.58% higher than the single-modal text and image sentiment recognition algorithms, respectively. On two public datasets, MVSA-Single and MVSA-Multi, the model in this paper achieves an accuracy of 79.48% and 77.79%, respectively, with F1 values of 75.64% and 74.33%, which is superior to the current mainstream models such as DR-Transformer. The feature-level fusion strategy further improves the model performance, and the accuracy is improved by 1.32% over the fused feature decision, reaching 89.56%, indicating that fully utilizing the inter-modal correlation and independence information has significant value for sentiment analysis. The simulation experiments verify the applicability of the model in a real multi-source social network environment, and the resulting positive sentiment value of 8636.6 is highly consistent with the theoretical value of 8636.1. These results show that cross-modal feature fusion can effectively solve the limitations of single-modal sentiment analysis, and through the complementation and enhancement of graphical and textual information, it can provide a more comprehensive and accurate understanding of social network users' emotions, and provide reliable technical support for opinion monitoring, personalized recommendation and user behavior analysis.

Acknowledgements

The team has won the third prize in the selection round of the “China College Students' 'Internet+' Innovation and Entrepreneurship Competition”.

References

[1] Alafwan, B., Siallagan, M., & Putro, U. S. (2023). Comments Analysis on Social Media: A Review. EAI Endorsed Transactions on Scalable Information Systems, 10(6).

[2] Vollero, A., Sardanelli, D., & Siano, A. (2023). Exploring the role of the Amazon effect on customer expectations: An analysis of user-generated content in consumer electronics retailing. Journal of Consumer Behaviour, 22(5), 1062-1073.
[3] Stefan, M. C., Andreiana, V. A., & Panagoret, I. (2017). THE ROLE OF SOCIAL NETWORKS IN THE EVOLUTION OF ONLINE SALES-STUDY CASE. Journal of Science & Arts, 17(4).
[4] Kim, Y., & Jang, A. (2021). A longitudinal study of sales promotion on social networking sites (SNS) in the lodging industry. Journal of Hospitality and Tourism Management, 48, 256-263.
[5] Toussaint, P. A., Renner, M., Lins, S., Thiebes, S., & Sunyaev, A. (2022). Direct-to-Consumer Genetic Testing on Social Media: Topic Modeling and Sentiment Analysis of YouTube Users' Comments. JMIR infodemiology, 2(2), e38749.
[6] Zhou, Q., Xu, Z., & Yen, N. Y. (2019). User sentiment analysis based on social network information and its application in consumer reconstruction intention. Computers in Human Behavior, 100, 177-183.
[7] Micu, A., Micu, A. E., Geru, M., & Lixandroiu, R. C. (2017). Analyzing user sentiment in social media: Implications for online marketing strategy. Psychology & Marketing, 34(12), 1094-1100.
[8] Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management, 48, 280-290.
[9] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN computer science, 3(1), 74.
[10] Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. Procedia Computer Science, 113, 65-72.
[11] Jabalameli, S., Xu, Y., & Shetty, S. (2022). Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: At the early stage of vaccination. International Journal of Disaster Risk Reduction, 80, 103204.
[12] Lyu, Y., Chow, J. C. C., & Hwang, J. J. (2020). Exploring public attitudes of child abuse in mainland China: A sentiment analysis of China's social media Weibo. Children and Youth Services Review, 116, 105250.
[13] Khalafat, M., Ja'far, S. A., Al-Sayyed, R., Eshtay, M., & Kobbaey, T. (2021). Violence detection over online social networks: An Arabic sentiment analysis approach. iJIM, 15(14), 91.
[14] Lee, S., Ma, S., Meng, J., Zhuang, J., & Peng, T. Q. (2022). Detecting sentiment toward emerging infectious diseases on social media: a validity evaluation of dictionary-based sentiment analysis. International Journal of Environmental Research and Public Health, 19(11), 6759.
[15] Hariguna, T., & Rachmawati, V. (2019). Community opinion sentiment analysis on social media using Naive Bayes algorithm methods. International Journal of Informatics and Information Systems, 2(1), 33-38.
[16] Wicaksono, F. A., & Romadhony, A. (2022). Sentiment analysis of university social media using support vector machine and logistic regression methods. Indonesian Journal on Computing (Indo-JC), 7(2), 15-24.
[17] Eldeeb, H., & Elshawi, R. (2024). Empowering Machine Learning with Scalable Feature Engineering and Interpretable AutoML. IEEE Transactions on Artificial Intelligence.
[18] Islam, M. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, M. M., & Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. Artificial Intelligence Review, 57(3), 62.
[19] Kaur, G., & Sharma, A. (2023). A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. Journal of big data, 10(1), 5.
[20] Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. EURASIP journal on wireless communications and networking, 2017, 1-12.
[21] Qinfu Xu,Yiwei Wei,Shaozu Yuan,Jie Wu,Leiquan Wang & Chunlei Wu. (2024). Learning emotional prompt features with multiple views for visual emotion analysis. Information Fusion,108,102366-.
[22] Raheem Mafas & Chong Yi Chien. (2024). E-Commerce Fake Reviews Detection Using LSTM with Word2Vec Embedding. Journal of computing and information technology,32(2),65-80.
[23] Liu Ning & Zhao Jianhua. (2022). A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text. Computational intelligence and neuroscience,2022,8726621-8726621.
[24] Ghotekar Rahul Krishnaji,Rout Minakhi & Shaw Kailash. (2023). Hybrid ResNet152-EML model for Geo-spatial image classification. International Journal of Information Technology,16(2),659-673.