Cross Modal Sentiment Analysis of Image Text Fusion Based on Bi LSTM and B-CNN

Yuan Fang[1*], Yi Wang[2]
[1]The College of Art and Media, Xianda College of Economics and Humanities Shanghai International Studies University Shanghai 200000, China
[2]EMC Information Technology R&D Shanghai Co., Ltd Shanghai 200000, China
E-mail: fangyuan_1122@126.com, wangyi_314159@163.com
[*]Corresponding author

Keywords: hybrid fusion strategy, image text fusion, cross-modal, sentiment analysis, multi-layer attention mechanism, CNN, BI-LSTM+ATT

Received: July 23, 2024

Due to the different modalities of data such as images and text, the difficulty of sentiment analysis increases. To achieve cross-modal sentiment analysis, the study firstly designs a cross-modal sentiment analysis method based on bi-directional long and short-term memory networks and bi-linear convolutional neural networks. At the same time, concepts such as image attributes are introduced in the experiment to detect irony in graphic and textual data. Finally, a hybrid strategy cross-modal sentiment analysis method is established in the experiment. After comparison, the proposed method has the highest subject working characteristic curve and PR, which are 5% and 3% higher than the comparative methods, respectively. The model has the lowest error take, with a minimum value of only 0.71%. The average F1 value and average accuracy reached 92.61% and 88.97%, respectively. When the validation set size is 400, the recognition time of the proposed method is 2.1 seconds. When iterating 50, the recognition time of this method is 0.9 seconds. In practical applications, the proposed method has accurately analyzed six types of graphic and textual content with different emotional tendencies. This method has the best detection results for both single graphic and cross-modal modes.

Povzetek: Večmodalna analiza sentimenta na podlagi Bi-LSTM in B-CNN učinkovito prepoznava čustva v vizualnih in tekstovnih vsebinah.

1 Introduction

In recent years, with the quick advancement of Internet technology in China, people's daily activities increasingly rely on the Internet. By December 2023, the Internet penetration rate in China has arrived 77.5%, and the amount of Internet users has reached 1.092 billion [1]. In addition, the amount of online video users is 1.067 billion, accounting for 97.7% of the total number of users. Internet users use video and other methods to record, thus generating massive multi-modal data. As the most basic way of communication, text is also the most important emotional carrier on the Internet. The research on sentiment analysis of texts has always been a focus of scholars' attention. Traditional sentiment analysis techniques are often designed for single-modal media such as text or images. Huang and Alias comprehensively analyzed the trend of the transformation of text sentiment analysis tasks in recent years and summarized the latest methods in the past 5 years. They believed that text sentiment analysis has shifted from keyword comparison based to deep learning algorithms [2]. Alwehaibi et al. combined Convolutional Neural Network (CNN) to train and test the modern standard Arabic and Arabic dialect data collected by Teitter. The outcomes denoted that the classification accuracy of Long Short-Term Memory (LSTM) and CNN models was 69.7%-88%, while the highest accuracy of the ensemble model was as high as 96.7% [3]. Although deep learning could significantly improve text recognition accuracy, there were issues with class imbalance and unlabeled corpus. Jiang et al. developed a new model that integrated K-means++, Synthetic Minority Oversampling Technique (SMOTE), CNN, and Bi-directional-LSTM (Bi-LSTM), which could adjust the data distribution of different emotional corpora and solve the problem of corpus imbalance [4]. Deep learning algorithms have become the mainstream trend in text sentiment analysis tasks, giving rise to various optimization algorithms, such as multi-layer classification model of Eahman et al. [5] and emerging metaheuristic optimization algorithm of Hosseinalipour et al [6].

Chinese users are increasingly inclined to communicate through images, emojis, and other means. Simple text sentiment analysis algorithms are gradually unable to meet the demand. Compared with the uni-modal data, multi-modal data has complex emotional information and universal relevance. These technologies are difficult to accurately identify and classify emotions contained in multi-modal data. Deng et al. applied an advanced uni-modal sentiment analysis model to process multi-modal data based on the integration idea and designed a multi-modal sentiment analysis method based on cross-modal attention and difference loss. The method employs cross-modal attention to achieve the integration of multi-modal information, while using difference loss to minimize the gap between image and text information. Experimental results showed that the model outperformed the baseline model on five publicly available datasets [7]. To take care of elderly users, mainstream social software such as WeChat and QQ have voice and video functions. The rise of short video platforms such as Tiktok has transformed the object of emotional analysis from uni-modal to multi-modal. This requires sentiment analysis algorithms to have a joint learning paradigm, namely context dependency, multi-modal interaction, and multi-task relevance [8]. The main goal of algorithms for sentiment analysis of images and text is to convolve and pool image features to extract features. Liao et al. proposed a graphic interactive graph neural network that utilizes text level graph neural networks to extract text features. They used pre-trained CNN to extract image features and constructed a graphic text interaction network. Finally, sentiment classification was implemented by combining the image text aggregation layer [9]. These models have effectively completed the task of sentiment analysis of graphic and textual information. However, they performed poorly on more complex and diverse video sentiment analysis.

The research focuses on how to effectively extract and jointly represent multi-modal features. The cross-modal contrastive learning proposed by Yang et al. successfully solves the heterogeneity problem caused by modal differences by integrating multiple contrastive learning algorithms and multi-modal data augmentation algorithms.

Table 1: Summary of related work.

Number	Author	Methods	Results	Deficiencies or research gaps
[2]	Hung L P, Alias S	An overview of the latest trends and challenges in text sentiment analysis to sentiment detection	The trend in text-based sentiment detection has shifted towards machine learning and deep learning algorithms	The real-time nature of the synthesis content is lacking
[3]	Alwehaibi A, Bikdash M, Albogmi M, et al	Integrated models for LSTM, CNN	The highest accuracy of the integrated model is 96.7% with the lowest loss value	There is still room for improvement in model performance, and the cost of model computation is large
[4]	Jiang W, Zhou K, Xiong C, et al	Integration of K-means++, SMOTE, CNN and Bi-LSTM models	Outperforming other models in text sentiment classification.	Integration complexity is too high
[5]	Rahman H, Tariq J, Masood M A, et al	Integration of decision trees, support vector machines and naïve bayesian models	Multi-layer model has 0.91 recall, 0.88 F1 value and feature optimization improves the efficiency of the model	Integration model complexity is high
[6]	Hosseinalipour A, Ghanbarzadeh R	Horse herd optimization algorithms	Better performance in sentiment analysis, improving the efficiency of the algorithm in terms of computational complexity	Data feature level optimization is not considered
[7]	Deng Y, Li Y, Xian S, et al	Multi-modal sentiment analysis methods based on cross-modal attention and disparity loss	Outperforms baseline methods on five datasets	Only sentiment accuracy analysis is considered, but not the underlying data feature processing.
[8]	Zhang Y, Jia A, Wang B, et al	Multi-modal, multitask interactive graph attention networks	Marginal rates of 1.88%, 5.37% and 0.19% for sentiment analysis and 1.99%,	High complexity of multi-task learning

			3.65% and 0.13% for sentiment recognition on the baseline datasets MELD, MEISD, and MSED respectively	
[9]	Liao W, Zeng B, Liu J, et al	Textual interactive graph neural networks	The method outperforms existing baseline models	Text features are extracted using only text-level graph neural networks
[10]	Yang S, Cui L, Wang L, et al	Cross-modal contrast learning integrating multiple contrast learning methods and multi-modal data augmentation	State-of-the-art performance is achieved on the three baseline datasets	The cross modal framework is too complex
[11]	Chen R, Zhou W, Li Y, et al	A video-based cross-modal assistive network	Higher classification accuracy for multi-modal sentiment analysis	Multi-modal integration, low analysis efficiency
[12]	Dong S, Fan X, Ma X	A multi-channel cross-modal feedback interaction model	Good performance in terms of accuracy and F1 scores	Higher complexity of model integration

The principle is to capture the complementarity of modal specific features by restricting modal features to different feature spaces [10]. On the basis of considering effective joint representation, Chen et al. further considered the shortcomings of single modal feature extraction and the problem of data redundancy in multi-modal fusion process. They proposed a cross-modal auxiliary network based on video. This network consisted of an audio feature map module and a cross-modal selection module. At the same time, a classifier composed of multiple image classification networks was introduced in the experiment to predict emotion polarity and emotion category [11]. Some scholars have also proposed the integration of external knowledge as a means of obtaining more accurate emotional information. Dong et al. put forth a multi-channel cross-modal feedback interaction model that incorporates knowledge graphs into multi-modal sentiment analysis. This model comprised a cross-modal feedback loop interaction module and an external knowledge module for the capture of potential information. During the training process, a self-feedback mechanism was adopted, and a global feature fusion module was utilized for the integration of multi-channel and multi-modal information [12]. The relevant work is summarized in Table 1.

In Table 1, existing multi-modal sentiment analysis mostly adopts integration schemes, and there is still room for improvement in performance when the model complexity is high. Meanwhile, the existing sentiment analysis mainly focuses on extracting high-level features of images and texts [13]. However, these methods ignore the underlying features of the image. At the same time, most methods use a uni-modal analysis to process image, speech, and text information. These lead to low classification accuracy in sentiment analysis. Cross-modal sentiment analysis (CMSA) is a type of multi-modal sentiment analysis. CMSA can fuse features from different modalities, such as images, speech, and text [14]. This method can establish connections between different modalities and extract relationships between image, speech, and text features. This effectively improves the classification accuracy of sentiment analysis methods. Therefore, a CMSA method is constructed in this study. In addition, an irony detection model is introduced in the experiment. It is hoped to provide reference for sentiment analysis in different modalities.

2 Methods and materials

2.1 Sentiment analysis method based on hybrid fusion strategy

Cross-modal feature fusion is the foundation of this method, which mainly includes model-based fusion and model independent fusion [15]. Among them, model-based fusion is mainly based on methods such as neural networks. These methods may reduce comprehension ability when conducting sentiment analysis due to the increase in modal types. The three main approaches to

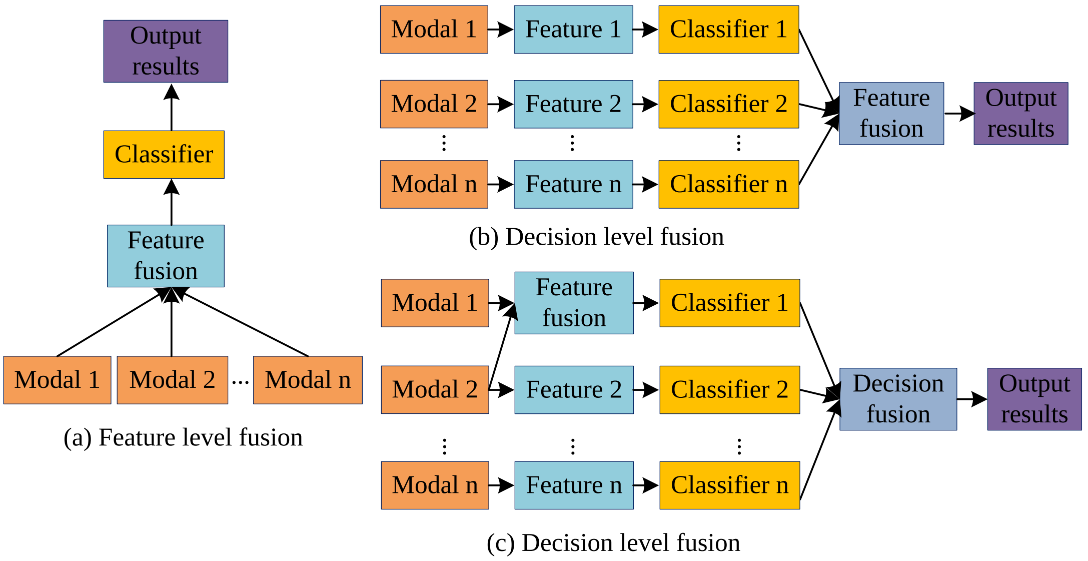


Figure 1: Model independent fusion.

model-independent fusion are feature-based early fusion, decision-based late fusion, and hybrid fusion. These methods can perform information exchange processing between different modal features. Therefore, these three methods can effectively perform cross-modal feature fusion and can be used for sentiment analysis. Therefore, a fusion method independent of the model is chosen in this study. Figure 1 is a flowchart of these three fusion methods.

In Figure 1, feature level fusion can concatenate features between different modalities [16]. Then, the fused features generated after concatenation are transformed into vectors for model training. This fusion occurs during the input stage, hence it is also known as early fusion. In the later stage, there is no need for complex training, and simple classification can be used to complete the training. However, this method is easily constrained by the results of different modal features. In addition, due to the influence of splicing methods, it is difficult to effectively utilize the feature information between different modals. Decision level fusion is mainly performed in the later stages of training, hence it is also known as late stage fusion [17]. Each modal can be trained separately without interfering with each other, thus having high robustness. However, this fusion method cannot fully understand the correlation between different modal information in feature level fusion. Therefore, this method is difficult to achieve feature fusion of different models. Hybrid fusion integrates two methods: feature level fusion and decision level fusion. This method can flexibly process different modal features during training. Therefore, a hybrid fusion method is used as the framework in the experiment to establish a CMSA model for image text fusion.

In the CMSA model of image text fusion, it is necessary to process image text information. CMSA mainly extracts features from different modalities and then performs fusion analysis. This process involves feature extraction, which has a significant impact on hybrid fusion. The main methods are deep learning techniques such as CNN and Transformer, as well as multi-task learning [18]. The existing CMSA methods can explore the information connections between different modals through feature alignment. In addition, it is necessary to utilize non-feature alignment to fuse information between different

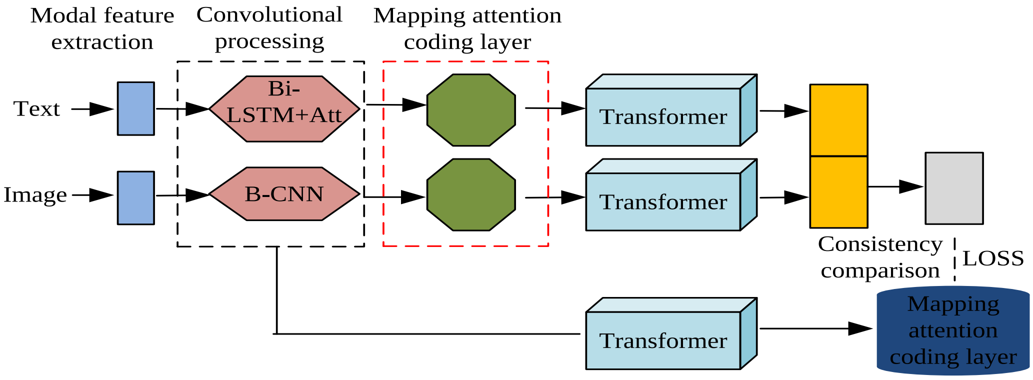


Figure 2: Structure of CMCC.

modalities. Therefore, a CMSA method, Cross Mapping and Consistency Comparison Model (CMCC), is proposed in this study. CMCC can map secondary modal information to primary modal information, which is beneficial for enriching the feature information of different modalities. Figure 2 shows the structure of CMCC.

CMCC mainly consists of four main structures. Firstly, modal feature extraction is used for primary feature processing of different modalities of text and images. The extraction of text features adopts a combination of Bi-LSTM and attention mechanism (Att), namely Bi-LSTM+Att [19]. The feature extraction of the image adopts bilinear CNN, namely B-CNN. The second step is to input the obtained text and image features into the Mapping Attention Coding Layer (MACL) for information exchange in different modalities. The next step is to use Transformer encoder for advanced feature extraction of different modalities. This can reduce the long-range dependencies of different modalities and help fuse non-aligned features. Finally, there is a comparison of consistency among different modalities. By establishing a consistency comparison structure and using a loss function for model training, the learning ability of CMCC is ultimately improved.

In CMCC, the first step is to extract features from text and image information. Text features are mainly represented by vectors of sentences and words. The existing methods for representing sentences and word vectors, such as One-Hot and Glove, have certain issues such as vocabulary gaps [20]. To better measure the semantic distance of different words, the Bi-LSTM+Att structure is used in the experiment. The extraction of text features adopts Bi-LSTM+Att. Figure 3 shows the specific extraction method.

Bi-LSTM+Att mainly has 5 layers structure including input layer and Bi-LSTM, etc. Firstly, the input layer is used for inputting textual data. Secondly, the embedding layer utilizes Transformer to embed text information into the vector space. The dimensions of this space are fixed. Formula (1) is the hidden layer state.

$$H_i^0 = W_{emb} x_i \qquad (1)$$

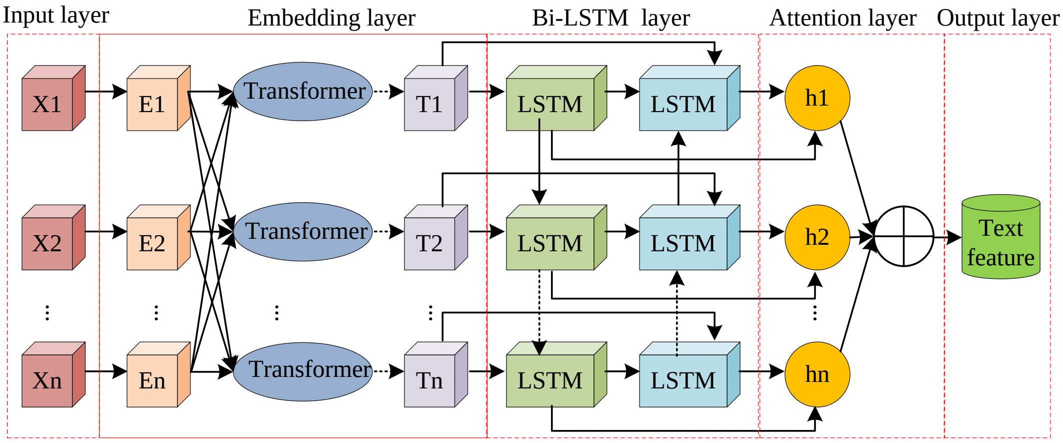


Figure 3: Structure of Bi-LSTM+Att.

In formula (1), i refers to the serial number. W_{emb} refers to the word embedding matrix. x refers to a word. H_i^0 is the hidden state of the 0th layer. Thus, the final output in formula (2) is obtained.

$$H_i^n = Transformer(H_i^{n-1}) \qquad (2)$$

In formula (2), n refers to the number of layers. The Bi-LSTM layer is capable of processing word vectors and obtaining advanced features, including both forward and backward directions. Formula (3) is the calculation of forward LSTM.

$$h_t^f = LSTM_f(h_{t-1}^f, [H_t; c_t^f]) \qquad (3)$$

In formula (3), $LSTM_f$ refers to the positive unit. h_t^f refers to the forward hidden state during t. H_t refers to output. c_t^f refers to positive output. [] means splicing behavior. Formula (4) is the calculation of inverse LSTM.

$$h_t^b = LSTM_b(h_{t-1}^b, [H_t; c_t^b]) \qquad (4)$$

In formula (4), $LSTM_b$ refers to the reverse unit. h_t^b refers to the reverse hidden state during t.

By weighting, the attention layer recombines the features output from the previous layer and obtains a weight vector. Then, through multiplication, sentence level feature vectors can be obtained. Formula (5) is the forward attention weight α_t^f for t.

$$\alpha_t^f = soft\max(w_{att} \tanh(W_c[H_t; c_t^f])) \qquad (5)$$

In formula (5), w_{att} and W_c refer to weights. $\tanh$ refers to the tangent function. Finally, there is the output layer, which is used for outputting vector sequences and ultimately obtaining text features. Formula (6) is the final positive eigenvector v_t^f in t.

$$v_t^f = \alpha_t^f [H_t; c_t^f] \quad (6)$$

Formula (7) is the final inverse eigenvector v_t^b for t.

$$v_t^b = \alpha_t^b [H_t; c_t^b] \quad (7)$$

The above v_t^f is concatenated with v_t^b to obtain the comprehensive text feature vector in formula (8).

$$v_t = [v_t^f; v_t^b] \quad (8)$$

Analyzing the emotional information contained in images is more challenging than simply identifying objects. This involves extracting features from different levels of abstraction. Therefore, the feature processing of images in sentiment analysis requires balancing different levels of features and integrating them. This is related to the final output of sentiment analysis methods. The methods for extracting image features include deep learning methods such as CNN and manually designed extractors such as SIFT. CNN can automatically extract deeper features from images. At the same time, CNN can comprehensively express emotional features. Therefore, in this study, a B-CNN is constructed based on CNN. Figure 4 shows the structure of B-CNN.

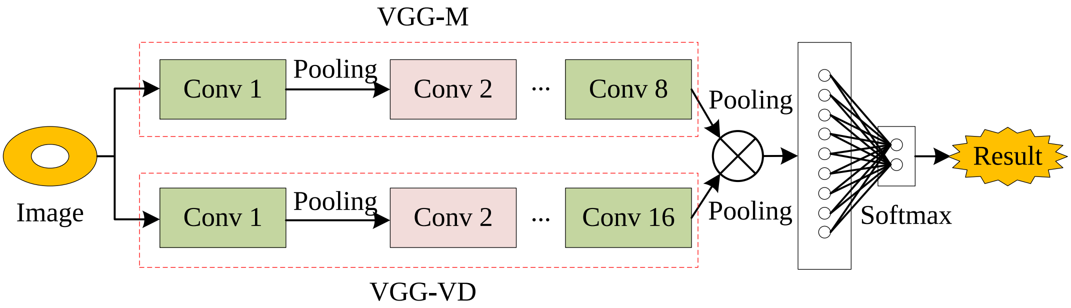


Figure 4: Structure of B-CNN.

B-CNN can obtain interactive information of different image features through bilinear pooling. Two types of CNNs are used in this structure for processing input data. The convolution kernel used in the experiment is 3×3. Each CNN uses VGG-Net for convolutional feature extraction. These two structures can simultaneously extract left and right features from input data. After feature fusion, these features can be better captured. Finally, feature output is achieved through a fully connected layer.

MACL is an improved version based on Transformer. MACL can use multi-head attention mechanism to judge the relationship between different modal information. This approach does not require modal feature alignment and can obtain more cross-modal emotional representations. The query, key, and value matrices of MACL are redefined in formula (9).

$$\begin{cases} Q_\alpha = X_\alpha W_{Q_\alpha} \\ K_\beta = X_\beta W_{K_\beta} \\ V_\beta = X_\beta W_{V_\beta} \end{cases} \quad (9)$$

In formula (9), Q_α refers to the query matrix. K_β stands for Key Matrix. V_β refers to the matrix of values. X refers to the input matrix. W refers to the linear transformation matrix. α and β refer to two different modes. Therefore, the process of mapping β to α in formula (10) can be obtained.

$$U_\alpha = soft\max(\frac{Q_\alpha K_\beta^T}{\sqrt{d_K}})V_\beta \quad (10)$$

In formula (10), $\sqrt{d_K}$ refers to the scaling factor used for scaling. T refers to location information. The dimensions of Q_α and U_α remain consistent. But the mapping of β is obtained in U_α. Residual connections are introduced in the experiment to prevent network degradation caused by multi-layer network connections. Figure 5 shows the structure of MACL.

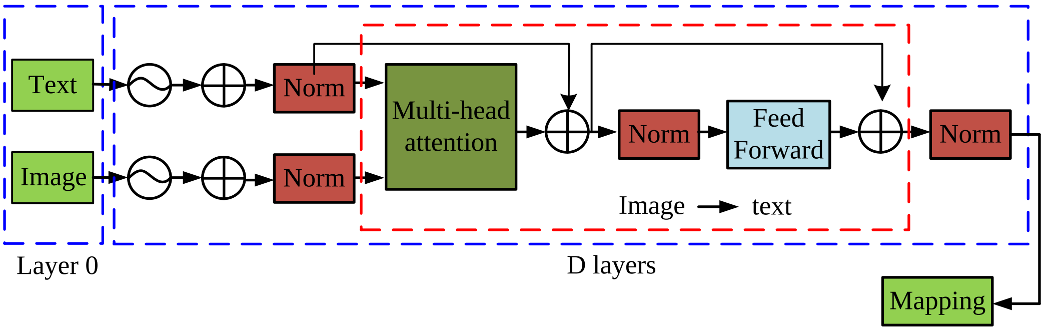


Figure 5: Schematic diagram of MACL structure.

In Figure 5, image and text information are used as model inputs. MACL consists of D layers, each of which contains modules for mapping images to text. The experiment includes both text and image modals, therefore an MACL module for text processing and an image processing module are set up. The MACL module can map image data to text data. Finally, the fused features are obtained through mapping. Through the processing of the MACL module, information heating interaction is achieved by mapping between different modules. This can solve the long-distance dependency problem.

2.2 Cross-modal irony detection method based on image text fusion

In social media, people also use rhetorical devices such as satire to express their emotions when expressing general opinions. At this point, there is a certain difference between the emotional meaning expressed by people and the surface meaning. Due to the ambiguity and polysemous nature of ironic language, it is often difficult to accurately capture its true intent by relying only on a single piece of textual data. In addition, the complexity of multi-modal data also increases the difficulty of irony detection, and multiple media such as images and sounds can present ironic expressions. Images can visually reflect the scene or emotion described in the text and provide important clues to understand the true intent of the text [21]. Based on this situation, images can be used as an important source of supplementary information, and the introduction of image information helps to analyze the speaker's true intention. To perform irony detection more accurately, the study makes full use of multi-modal data, combines multiple information sources such as text and images for comprehensive analysis, and proposes a cross-modal irony detection method based on graphic fusion.

The proposed cross-modal irony detection method for graphic fusion includes three types of modal features, namely text, image level image attributes. Image attributes are introduced in the experiment to detect irony in graphic and textual data. Image attributes can help people understand the content, features, and states in images. Image attribute features as text features and image features as an auxiliary, the research will be used as as a connecting bridge between the other two features, through the reasoning of image attributes of the image of the scene, emotional information. Formula (11) is the extracted image attribute vector V_{attr}.

$$V_{attr} = \sum_{i=1}^{5} A_i \cdot E(a_i) \tag{11}$$

In formula (11), $E(a_i)$ refers to the image attribute vector before extraction. A_i stands for weight. Before performing feature extraction, image datasets need to be processed through data augmentation. This can improve the pan Chinese ability of sentiment analysis models, reduce overfitting, and enhance robustness. The commonly used methods are

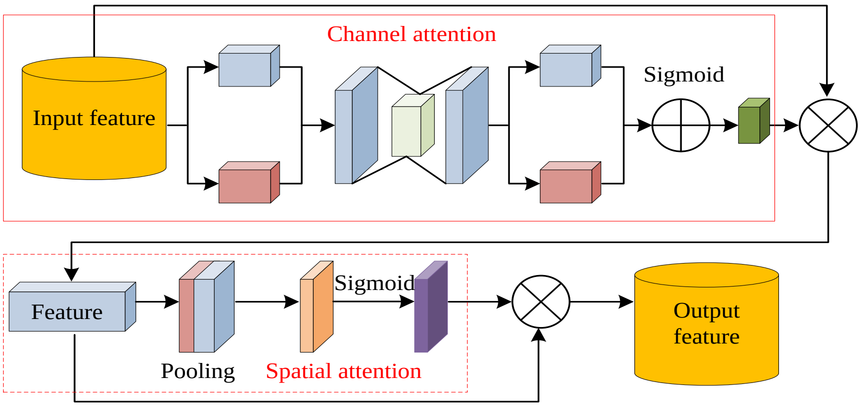


Figure 6: Structure of CBAM.

rotation, cropping, etc. OpenCV is used in the experiment to enhance image data, involving image cropping, inversion, and color space conversion [22]. Image inversion is the process of inverting the brightness value of each pixel of an image. Each pixel in an image has a value between 0 and 255, and the operation of inverting is to subtract the value of the current pixel from 255. Image inversion changes darker pixels to brighter and brighter pixels to darker, showing the features and details of the image by changing the brightness and contrast of the image. Image cropping is the process of selecting a rectangular area from the original image as a new image, which extracts the part of the image of interest and removes irrelevant background information. Colour space conversion is to convert an image from one colour space to another, using the cvtColor function in OpenCV to accomplish this operation.

After the above image enhancement processing, image features are extracted in the experiment. Image features can be used to determine the presence of unreasonable scenes or elements. The basic method used in the experiment is ResNet50 with the addition of Convolutional Block Attention Module (CBAM) [23-24]. Figure 6 shows the structure of CBAM.

CBAM includes spatial attention and channel attention. First is to input the intermediate feature map into CBAM. Then, the weighted intermediate feature map is obtained through the channel attention mechanism module. These images are then input into the spatial attention mechanism module to obtain spatial attention. Finally, the two modules are serialized and the final image features are outputted. CBAM can weight the dimensions and channels of images to enhance feature learning and improve the effectiveness of feature extraction [25-26]. CBAM will not significantly increase the number of parameters and computational difficulty, and has strong operability. Formula (12) is the extracted text feature V_{image}.

$$V_{image} = [F_1(X), F_2(X), ..., F_n(X)] \quad (12)$$

In formula (12), $F(X)$ refers to the output of the residual block. In cross-modal irony detection methods, feature extraction of the obtained text information is also required. Bi-LSTM+Att is used in the experiment to improve cross-modal representation. Formula (13) is the extracted text feature V_{text}.

$$V_{text} = LayerNorm(V_t) \quad (13)$$

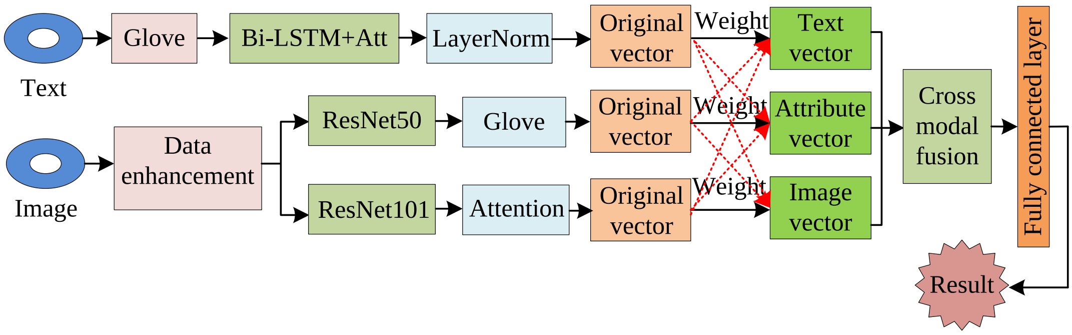


Figure 7: Cross-modal irony detection method.

In formula (13), $LayerNorm$ means normalization processing. V_t refers to the feature vector before processing. The combination of image attributes and graphic features can improve the robustness of CMSA methods. The image attributes and graphic features are obtained through the above method. However, the dimensions of these data are not uniform, making it difficult to capture the connections between different modalities. Therefore, after processing the graphic and textual features, cross-modal fusion is used in the experiment to process these data. Formula (14) is the fused feature vector V_{fused}.

$$V_{fused} = \sum_{m\{text,image,attr\}} \alpha_m V_m^{'} \quad (14)$$

In formula (14), m refers to different modes. $V_m^{'}$ refers to the processed feature vector. $attr$ refers to image attributes. The threshold node for the above method is 0.5. When the value is greater than this, it indicates that the result is judged as ironic. Less than this value is considered non-ironic. After calculation, the cross-modal irony detection method shown in Figure 7 is established in the experiment.

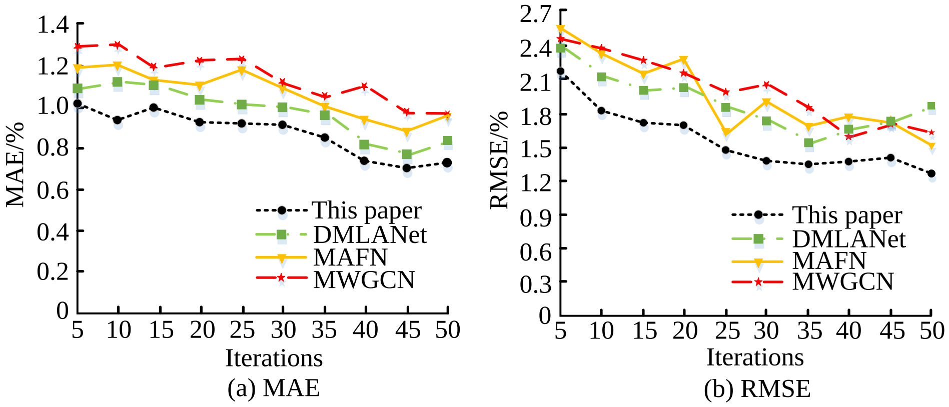
In this cross-modal irony detection method, text feature extraction and image data augmentation are first performed. For the processing of text modality, Bi-LSTM+Att processes the original vector of the text to obtain a weighted vector. For the processing of image modalities, ResNet50 and ResNet101 are used to process image features and image attribute modalities, respectively [27-28]. After weighted balancing, the image attribute vector and image vector are obtained. Then, the above three modalities are fused and passed through a fully connected layer to obtain the final output result.

On the basis of the above CMSA method based on image text fusion, an sentiment analysis system based on a hybrid fusion strategy is established in the experiment. This system is capable of analyzing graphic and textual information independently, as well as conducting CMSA. The system includes three modules: input, data processing, and result display. The graphic and textual data of people on social media is processed by the data processing module after passing through the input module. Polarity and confidence can be displayed in the result display module. Polarity includes both positive and negative emotional tendencies [29].

3 Results

3.1 Performance comparison of sentiment analysis methods based on hybrid fusion strategy

The proposed method was validated in the experiment. The operating system used for the experiment was Windows 10, the central processor was Intel(R) Core(TM) i9-9900X CPU@3.50GHz with 64.00 GB of RAM, the graphics processor was GeForce GTX 2080Ti with 32G, the algorithmic implementation language was Python 3.8, and the deep learning framework was Pytorch 1.7. The verification experiment was conducted in the Keras library written in Python. The number of training sessions was 50. The batch size of the dataset was 400. The model learning rate was 0.001. The text length was 60. Dropout was used to prevent over fitting phenomena with a parameter of 0.5. The number of dimensions in the fusion model were all 300. Pre-training was completed in Word2Vec with 100 dimensions. The dataset used for validation experiments in the experiment was MVSA for CMSA. The MVSA dataset combines information from both image and text modalities, which can help the model capture the emotions expressed by the user more comprehensively. Moreover, it contains annotated information with a wide range of sources and diversity. This dataset has two subsets, corresponding to 5129 and 19600 pairs of image and text data. 80% of the image and text data pairs in MVSA were selected as the training set. The remaining image and text data generated a validation set for the collection. The comparison methods selected in the experiment included Deep Multi-level Attention Network (DMLANet) [30], Multi-level Attention Fusion Network (MAFN) [31], and Multi-weight Graph Convolutional Network (MWGCN) [32]. These three models are existing state-of-the-art models that can effectively validate the study.

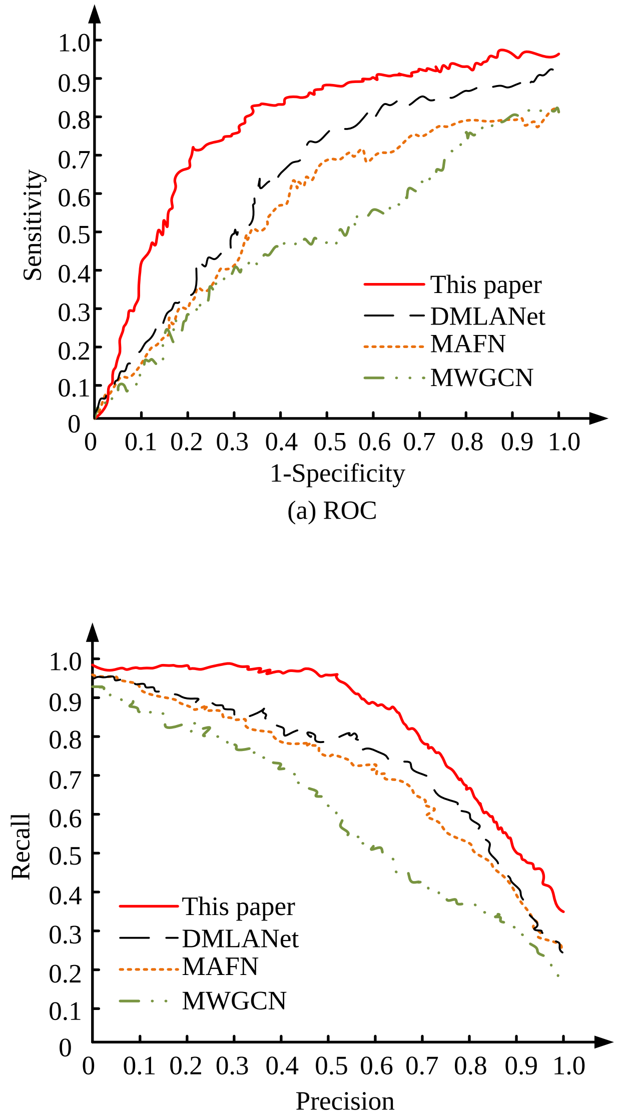


(a) MAE (b) RMSE

Figure 8: Comparison of MAE and RMSE.

The performance comparison of models usually involves comparing different indicators. Different indicators imply different model functionalities. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are common comparison metrics. MAE can reduce the possibility of errors offsetting each other. RMSE has high sensitivity and can effectively reflect the deviation between the true value and the measurement result. Both of these indicators can to some extent reveal the accuracy of the model. Figures 8 (a) and (b) correspond to the MAE and RMSE comparison results of DMLANet, MAFN, MWGCN, and the proposed method, respectively. Among all methods, the proposed method had the lowest MAE and RMSE. After 50 iterations, the MAE and RMSE of this method were 0.71% and 1.23%, respectively. The MAE of other methods was higher than 0.80%, and the RMSE was higher than 1.50%. Therefore, the proposed method this time had higher performance.

The above comparison of indicators cannot fully reflect the application effect of the model. Therefore, Receiver Operating Characteristic Curve (ROC) and PR curve



(a) ROC

(b) PR

Figure 9: Comparison of ROC and PR.

were introduced in the experiment for method comparison. ROC analysis can provide neutral decision results without being affected by factors such as cost. The PR curve can serve as a supplement to ROC analysis and also measure the classification ability of the model. Figures 9 (a) and (b) correspond to the ROC and PR comparison results of DMLANet, MAFN, MWGCN, and the proposed method, respectively. Among all methods, the proposed method had the highest ROC and PR. The ROC of this method improved by 5.8%, 6.2%, and 7.1% compared to DMLANet, MAFN, and MWGCN, respectively. The PR results of this method improved by 3.9%, 4.2%, and 5.4% compared to DMLANet, MAFN, and MWGCN, respectively. These results also confirmed that the proposed method was superior.

The results of the statistical analyses of different model performance evaluation metrics are shown in Table 2, where 'ROC-AUC' denotes the ROC-Area under Curve (AUC) and 'PR-AUC' denotes the PR-Area under Curve (AUC). In Table 2, the performance of the designed model was improved on MAE, RMSE, ROC, PR, and Accuracy, F1, and the difference in performance enhancement was statistically significant ($P<0.05$) when compared with other models.

Four types of emotional indicators were selected in Figure 11 for the detection of classification ability, including anger, happiness, sadness, and neutrality. The red data represents the result of classification errors. Figures 11 (a) to 11 (d) correspond to the classification results of MWGCN, MAFN, DMLANet, and the proposed method, respectively. The proposed method this time could accurately identify the highest proportion of data. In the classification of anger, happiness, sadness, and neutral expressions, the incorrect recognition results for both anger and happiness expressions were 3 times. The misidentification results for sad, neutral, and neutral expressions were all twice. The error recognition results of other methods were higher than this research method. Therefore, the proposed method had superior recognition and classification capabilities.

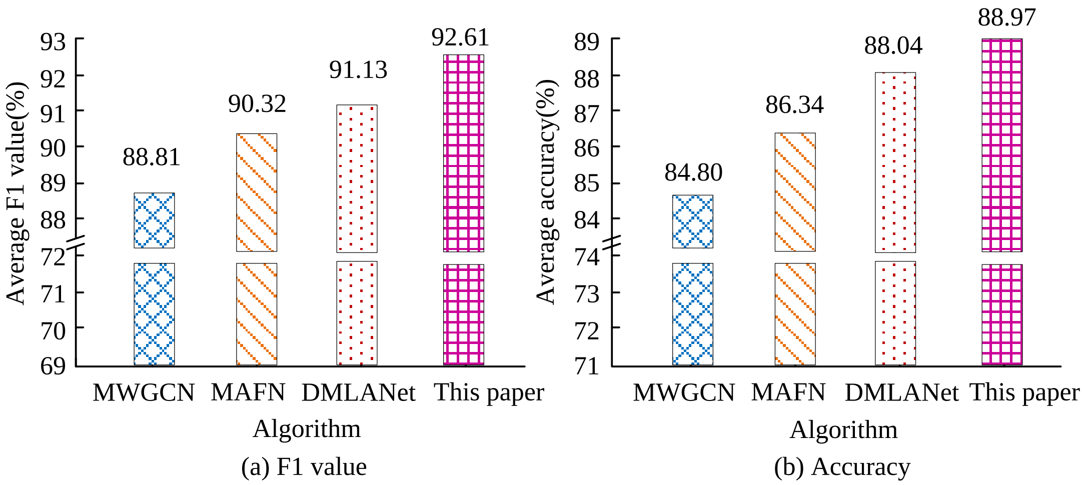


Figure 10: Average F1 value and accuracy.

Table 2: Statistical analysis of performance evaluation indicators of different models.

Evaluating index	**Model**				**P**
	DMLANet	**MAFN**	**MWGCN**	**This Paper**	
MAE(%)	0.915±0.061	1.015±0.081	1.235±0.095	0.711±0.225	0.038
RMSE(%)	1.805±0.271	1.511±0.062	1.652±0.144	1.231+0.135	0.033
ROC-AUC	0.861±0.166	0.763±0.179	0.694±0.231	0.941±0.036	0.026
PR-AUC	0.846±0.067	0.793±0.197	0.699±0.207	0.953±0.123	0.028
Accuracy(%)	88.81±1.06	90.32±0.656	91.13±0.106	92.61±0.186	0.038
F1	84.80±0.03	86.34±0.163	88.04±0.613	88.97±0.136	0.016

In the verification of the application effect of methods, time is also a factor that affects the application effect of the model. In practical applications, it is necessary to reduce the running time and save time costs while ensuring the performance of the model. Figures 12 (a) and 12 (b) correspond to the recognition time of the model under different validation set sizes and iteration times, respectively. In Figure 12 (a), increasing the size of the validation set significantly improved the recognition time of each method. When the validation set size was 400, the recognition times of MWGCN, MAFN, DMLANet, and the proposed method were 4.9s, 3.5s, 2.7s, and 2.1s, respectively. In Figure 12 (b), the increase in iterations had a greater impact on model performance, and the two were inversely proportional. When iterating 50, the recognition times of MWGCN, MAFN, DMLANet, and the proposed method were 2.3s, 2.0s, 1.8s, and 0.9s, respectively. Therefore, the proposed method had superior recognition efficiency.

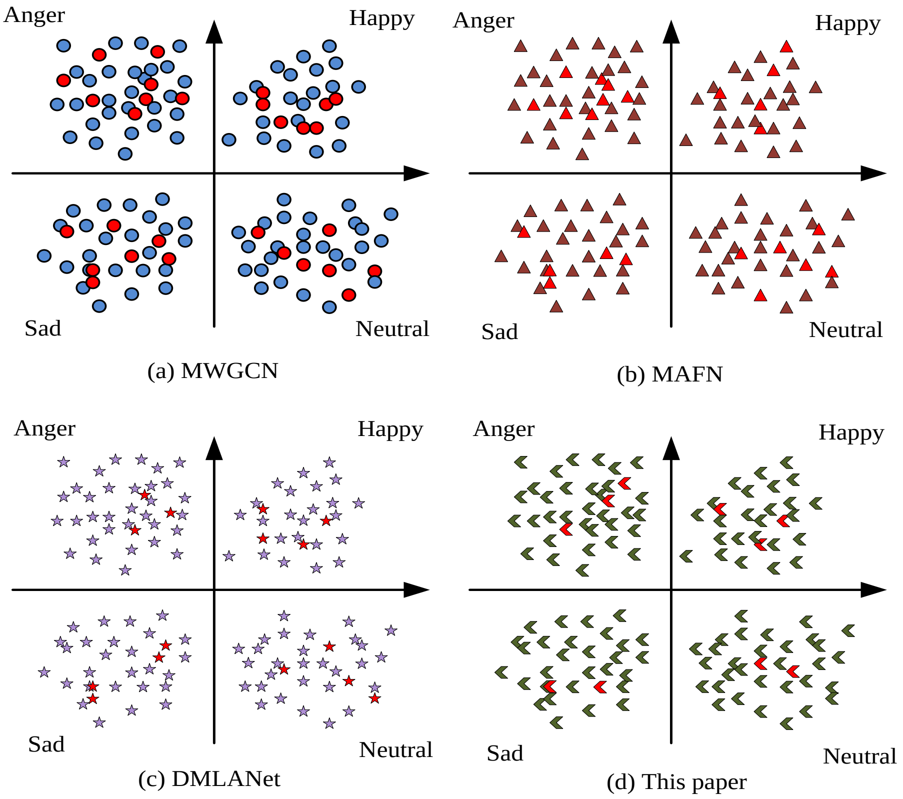
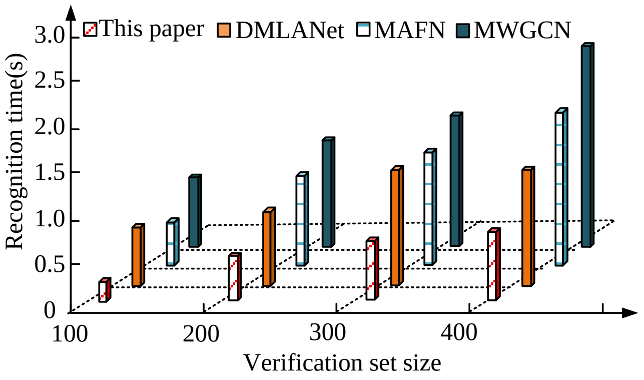
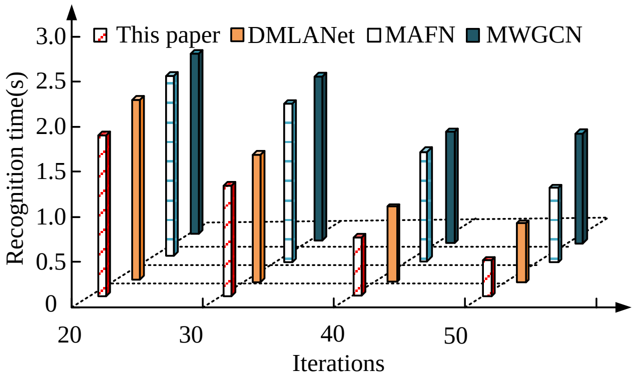


Figure 11: Comparison of recognition and classification performance.



(a) Recognition time under different validation set sizes



(b) Recognition time under different iterations

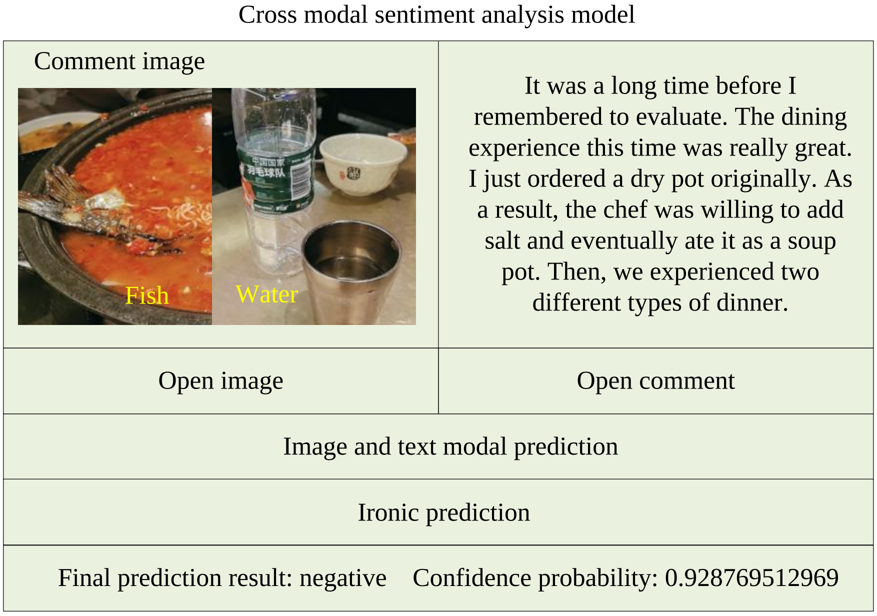
Figure 12: Recognition efficiency.

Table 3: Sentiment analysis results.

Image content	Text content	MWGCN	MAFN	DMLANet	Proposed method
Case 1	What a bustling and bustling scene outside the window. The air inside the window is so fresh, clean and tidy. I really admire it.	Negative	Negative	Neutral	Negative
Case 2	The service attitude needs to be improved.	Negative	Negative	Negative	Negative
Case 3	The dishes are very fresh. I will come again next time.	Positive	Positive	Positive	Positive
Case 4	Thick smoke in the air.	Neutral	Negative	Negative	Negative
Case 5	Partying with cousins!	Neutral	Positive	Positive	Positive
Case 6	I want to go home and see my dog.	Neutral	Positive	Positive	Neutral

3.2 Analysis of the practical application effect of sentiment analysis method based on hybrid fusion strategy

The hybrid strategy-based sentiment analysis method mentioned above showed superior performance in the comparison of various indicators. But this method still needs to be applied in practical sentiment analysis scenarios. Six sample cases were selected in the experiment for sentiment analysis using this method. Table 3 presents the analysis results of various methods. In six cases, the proposed method had accurate analysis results for graphic and textual content with different emotional tendencies. The analysis results of MWGCN, MAFN, and DMLANet in each case had certain errors. Therefore, the proposed method could effectively handle graphic and textual information processing under different emotional colors. This method had better sentiment analysis ability.


Cross modal sentiment analysis model
Comment image
Fish
Water
Open image
Open comment
Image and text modal prediction
Ironic prediction

It was a long time before I remembered to evaluate. The dining experience this time was really great. I just ordered a dry pot originally. As a result, the chef was willing to add salt and eventually ate it as a soup pot. Then, we experienced two different types of dinner.

Final prediction result: negative Confidence probability: 0.928769512969

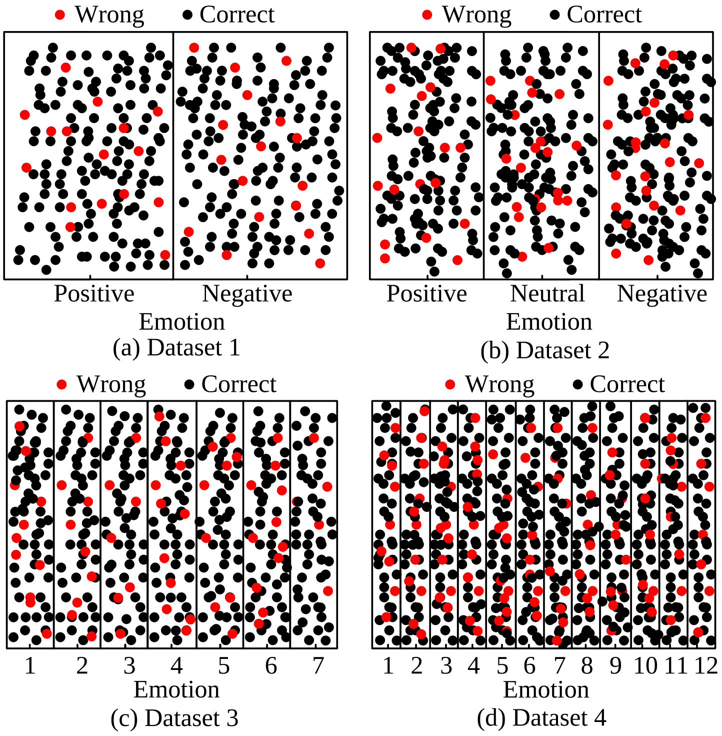
Figure 13: Cross-modal sentiment analysis.

In addition, the proposed method could also be used for irony detection. Therefore, in Figure 13, the detection results of this method in ironic image and text information were further analyzed. The case used in the experiment was the dining evaluation of a restaurant. The commentator's text information contains the word ‘great’. However, based on the visual and textual information, the reviewer discussed that the fish was very salty and added a lot of water, ultimately resulting in a soup pot. Therefore, based on the comprehensive image and text information, the reviewer used irony to evaluate the dining experience. The proposed method could accurately detect the emotional tendency of irony, with an accuracy rate of 0.9288. Therefore, the proposed method can be effectively used for irony detection.

The research method can be applied to the image text information detection of single mode and cross mode. Therefore, in Table 4, this method was compared between uni-modal and cross-modal methods in MVSA and Twitter datasets. The Twitter dataset contained a large amount of image text fusion data. The image and text information with a lot of irony was suitable for the irony detection of the proposed method. The indicators for comparison were accuracy and F1 value. From the table, the research method had the highest accuracy and F1 value results in both MVSA and Twitter. In the MVSA dataset, the accuracy of this method corresponded to 75.28%, 84.59%, and 86.34% for single and cross-modal images. The F1 values of this method corresponded to 74.51%, 82.69%, and 86.77%. In the Twitter dataset, the accuracy of this method corresponded to 73.02%, 82.05%, and 83.75% for a single image and text modal and cross-modal. The F1 values of this method corresponded to 72.27%, 80.21%, and 84.17%. Whether it was a single image and text modal or cross-modal, this method had the best results. These results further confirmed the superiority of this method.

Table 4: Detection results under different modals.

Modal	Model	MVSA		Twitter	
		Accuracy/%	F1/%	Accuracy/%	F1/%
Image	MWGCN	72.61	71.33	70.43	69.19
	MWGCN	72.35	71.65	70.18	69.50
	DMLANet	73.46	72.49	71.26	70.32
	Research method	75.28	74.51	73.02	72.27
Text	MWGCN	82.65	79.66	80.17	77.27
	MWGCN	83.01	80.47	80.52	78.06
	DMLANet	83.77	82.51	81.26	80.03
	Research method	84.59	82.69	82.05	80.21
Cross-modal	MWGCN	82.54	77.46	80.06	75.14
	MWGCN	82.97	78.03	80.48	75.69
	DMLANet	84.63	80.21	82.09	77.80
	Research method	86.34	86.77	83.75	84.17



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

Figure 14: Emotion classification under different dimensions.

Figure 14 shows the classification results of the proposed research method in different dimensions. Figure 14 (a) shows the classification results of positive and negative emotions, which belong to 2D data. Figure 14 (b) shows the classification results of positive, neutral, and negative emotions, which belong to 3D data. In Figure 14 (c), 1-7 refers to emotions such as joy, anger, worry, thought, sadness, fear, and shock, which belong to 7D data. In Figure 14 (d), 1-12 represent emotions such as happiness, sadness, excitement, worry, disgust, surprise, deep love, happiness, neutrality, anger, boredom, and emptiness, respectively, belonging to the 12D data. From the figure, the proposed method had the best classification effect on 2D data and the worst classification effect on 12D data. This indicated that as the dimensionality increased, the classification performance of the proposed method this time decreased. When the dimension was 7, this method still showed good classification performance.

4 Discussion

The development of information technology has driven the rise of multi-modal data content, changed the way people express and communicate, enriched the dimension of information dissemination, and brought development opportunities for the field of sentiment analysis. Multi-modal sentiment analysis is to extract and parse sentiment information in different data forms such as text, image, and speech. Compared with uni-modal sentiment analysis, multi-modal sentiment analysis is able to present emotions in a more comprehensive and three-dimensional way. Different modal data can in turn complement each other's information, make up for the misjudgment caused by the lack of uni-modal information or noise interference, and improve the accuracy and robustness of emotion recognition.

To achieve accurate and efficient multi-modal sentiment analysis, a large number of scholars have carried out analysis and research on this. Researchers mostly use uni-modal sentiment analysis model integration to achieve multi-modal sentiment analysis. In literature [4], Jiang W et al. constructed an integrated model of K-means++, SMOTE, CNN and Bi-LSTM. In literature [5], Rahman H et al. constructed an integrated model of Curse Tree, Support Vector Machines, and Naïve Bayes, and in terms of accuracy, recall and F1 value, the integrated model achieved a better performance of 96.7%, 0.91, 0.88 respectively, but the integrated model had too many integration objects and the complexity of the integration strategy also increased the complexity and computational cost of the model. Up to this point, in literature [6], Hosseinalipour A et al. introduced intelligent optimization algorithms into the field of multi-modal sentiment analysis. However, multi-modal sentiment analysis still suffers from the shortcomings of a complex framework, ignoring the underlying features of the image, and failing to tap into the text irony information.

In this regard, the research followed the integration idea and designed a multi-modal sentiment analysis model based on Bi-LSTM and B-CNN, which decomposed and mapped the image and text and then spliced and fused them, and the cross-modal fusion solved the shortcomings of the traditional integration model that payed insufficient attention to local features. Meanwhile, to intend the multi-modal irony information, the study designed multi-modal irony detection based on three feature extraction modules, and the experimental results showed that the method had the lowest error value, the lowest value was only 0.71%. The average F1 value and the average accuracy rate reached 92.61% and 88.97%, respectively. Moreover, the method had high recognition efficiency and could accurately detect this ironic emotional tendency in practical applications with an accuracy rate of 0.9288. In comprehensive analysis, the research-designed method using bilinear feature fusion not only improved the interaction efficiency of multi-modal data, but also made it easier to obtain the complete and locally important information, and thus improved the performance index significantly. The splicing fusion strategy compensated for the neglect of local features by simple integration.

However, this bilinear fusion strategy still has some cited limitations. On the one hand, it is a serious technical challenge to achieve the best effect of the splice fusion of two low-dimensional mapping matrices, and the process involves multiple parameters and steps, which requires several fine adjustments and verifications. On the other hand, the generalization performance of the method has yet to be improved and validated, and the variability between different datasets as well as different tasks leads to the challenge of the model's generalization ability. How to ensure the fusion advantage in the face of complex or noisy data still needs to be further investigated.

5 Conclusion

The popularity of social networking sites has diversified the ways people express their emotions. People can express their emotional tendencies through images, text, voice, and video. However, diverse expressions may lead to biases in people's understanding of emotions. Therefore, it is necessary to establish a more accurate sentiment analysis model. This is beneficial for industries such as catering and hotels to accurately judge customer evaluations. At the same time, this is also conducive to regulating the online environment and guiding public opinion. Due to the different modalities of data such as images and text, the difficulty of sentiment analysis increases. Therefore, a CMSA method based on image text fusion was established in this study. In the experiment, Bi-LSTM+Att was used for text feature extraction and B-CNN was used for image feature extraction. At the same time, people's expressions have a certain emotional tendency. People may choose to express their emotions through irony. Therefore, based on the above CMSA method, concepts such as image attributes and CBAM were introduced in the experiment to detect irony in graphic and textual data. Based on the above content, a hybrid strategy CMSA method was established in the experiment. After comparison, the proposed research method had the highest subject working characteristic curve and PR, which were 5% and 3% higher than the comparative methods, respectively.

The average accuracy and F1 value of this method were the highest among all methods. The incorrect recognition results for both angry and happy expressions using this method were 3 times. Its incorrect recognition results for sad, neutral, and neutral expressions were all 2 times. When the validation set size was 400, the recognition time of the proposed research method was 2.1 seconds. When iterating 50, the recognition time of this method was 0.9 seconds. These results confirmed that the proposed method had superior recognition efficiency. In practical applications, the proposed method accurately analyzed six types of images and text content with different emotional tendencies. This method could accurately detect image and text information with ironic meaning, with an accuracy of 0.9288. In the MVSA and Twitter datasets, this method had the highest accuracy and F1 value for both single graphic and cross-modal modes. When the dimension was 7, this method still showed good classification performance. Whether it was a single image text modal or cross-modality, this method had the best results. These results further confirmed the superiority of this method. But this method still had shortcomings. For example, the experiment mainly involved cross-modal processing of images and text, neglecting the connection with modal such as speech and video. At the same time, CMSA methods also need to consider the influence of language. In the future, a cross-cultural sentiment analysis model can be established.

References

[1] X. Ma, "Internet use and income gaps between rural and urban residents in China," Journal of the Asia Pacific Economy, vol. 29, no. 2, pp. 789-809, 2024. https://doi.org/10.1080/13547860.2022.2054133

[2] L. P. Hung, and S. Alias, "Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 27, no. 1, pp. 84-95, 2023. https://doi.org/10.20965/jaciii.2023.p0084

[3] A. Alwehaibi, M. Bikdash, M. Albogmi and K. Roy, "A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 8, pp. 6140-6149, 2022 https://doi.org/10.1016/j.jksuci.2021.07.011

[4] W. Jiang, K. Zhou, C. Xiong, G. Du, C. Ou and J. Zhang, "KSCB: A novel unsupervised method for text sentiment analysis," Applied Intelligence, vol. 53, no. 1, pp. 301-311, 2023. https://doi.org/10.1007/s10489-022-03389-4

[5] H. Rahman, J. Tariq and M. A. Masood, "Multi-tier sentiment analysis of social media text using supervised machine learning," Computers, Materials &amp; Continua, vol. 74, no. 31, pp. 5527-5543. 2023. https://doi.org/10.32604/cmc.2023.033190

[6] A. Hosseinalipour, and R. Ghanbarzadeh, "A novel metaheuristic optimisation approach for text sentiment analysis," International Journal of Machine Learning and Cybernetics, vol. 14, no. 3, pp. 889-909, 2022. https://doi.org/10.1007/s13042-022-01670-z

[7] Y. Deng, Y. Li, S. Xian, L. Li, and H. Qiu, "MuAL: Enhancing multimodal sentiment analysis with cross-modal attention and difference loss," International Journal of Multimedia Information Retrieval, vol. 13, no. 3, pp. 31-62, 2024. https://doi.org/10.1007/s13735-024-00340-w

[8] Y. Zhang, A. Jia, and B. Wang, "M3GAT: A multi-modal, multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition," ACM Transactions on Information Systems, vol. 42, no. 1, pp. 1-32, 2023. https://doi.org/10.1145/3593583

[9] W. Liao, B. Zeng, J. Liu, P. Wei, and J. Fang, "Image-text interaction graph neural network for image-text sentiment analysis," Applied Intelligence, vol. 52, no. 10, pp. 11184-11198, 2022. https://doi.org/10.1007/s10489-021-02936-9

[10] S. Yang, L. Cui, L. Wang, and T. Wang, "Cross-modal contrastive learning for multimodal sentiment recognition," Applied Intelligence, vol. 54, no. 5, pp. 4260-4276, 2024. https://doi.org/10.1007/s10489-024-05355-8

[11] R. Chen, W. Zhou, Y. Li, and H. Zhou, "Video-based cross-modal auxiliary network for multimodal sentiment analysis," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32 no. 12, pp. 8703-8716, 2022. https://doi.org/10.1109/TCSVT.2022.3197420

[12] S. Dong, X. Fan, and X. Ma, "Multichannel multimodal emotion analysis of cross-modal feedback interactions based on knowledge graph," Neural Processing Letters, vol. 56, no. 3, pp. 1-17, 2024. https://doi.org/10.1007/s11063-024-11641-w

[13] P. Huang, S. Li, S. Li, Z. Liu, C. Zhang, Z. Z. Zhang, and Z. Liu, "The role of emotional sensitivity to missed opportunity and grey matter volume of thalamus in risk-taking behaviour," International Journal of Psychology, vol. 58, no. 4, pp. 360-367, 2023. https://doi.org/10.1002/ijop.12906

[14] H. Cheng, Z. Yang, X. Zhang, and Y. Yang, "Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3149-3163, 2023. https://doi.org/10.1109/TAFFC.2023.3265653

[15] L. He, S. Liu, R. An, Y. Zhuo, and J. Tao, "An end-to-end framework based on vision-language fusion for remote sensing cross-modal text-image retrieval," Mathematics, vol. 11, no. 10, pp. 2279-2295, 2023. https://doi.org/10.3390/math11102279

[16] Y. Zhan, J. Liu, and L. Ou-Yang, "scMIC: A deep multi-level information fusion framework for clustering single-cell multi-omics data," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 12, pp. 6121-6132, 2023. https://doi.org/10.1109/JBHI.2023.3317272

[17] Q. Zhang, H. Zhang, K. Zhou, and L. Zhang, "Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition," Tsinghua Science and Technology, vol. 28, no. 4, pp. 673-685, 2023. https://doi.org/10.26599/TST.2022.9010038

[18] P. Preethi, and H. R. Mamatha, "Region-based convolutional neural network for segmenting text in epigraphical images," Artificial Intelligence and Applications, vol. 1, no. 2, pp. 119-127, 2023. https://doi.org/10.47852/bonviewAIA2202293

[19] F. Cheng, L. Peng, H.Zhu, C. Zhou, Y. Dai, and T. Peng, "A defect data compensation model for infrared thermal imaging based on bi-lstm with attention mechanism," JOM, vol. 76, no. 6, pp. 3028-3038. 2024. https://doi.org/10.1007/s11837-024-06408-6

[20] H. Eskandari, M. Imani, and M. P. Moghaddam, "Best-tree wavelet packet transform bidirectional GRU for short-term load forecasting," The Journal of Supercomputing, vol. 79, no. 12, pp. 13545-13577, 2023. https://doi.org/DOI:10.1007/s11227-023-05193-4

[21] D. Tomás, R. Ortega-Bueno, G. Zhang, P. Rosso, and R. Schifanella, "Transformer-based models for multimodal irony detection," Journal of Ambient Intelligence and Humanized Computing., vol. 14, no. 6, pp. 7399-7410, 2023. https://doi.org/10.1007/s12652-022-04447-y

[22] V. D. Volokitin, E. P. Vasiliev, E. A. Kozinov, V. D. Kustikova, A. V. Liniov, and Y. A. Rodimkov, "Improved vectorization of opencv algorithms for RISC-V CPUs," Lobachevskii Journal of Mathematics, vol. 45, no. 1, pp. 130-142, 2024. https://doi.org/10.1134/S1995080224010530

[23] P. Bhuyan, P. K. Singh, and S. K. Das, "Res4net-cbam: a deep cnn with convolution block attention module for tea leaf disease diagnosis," Multimedia Tools and Applications, vol. 83, no. 16, pp. 48925-48947, 2024. https://doi.org/10.1007/s11042-023-17472-6

[24] X. Yang, Q. Zhang, S. Wang, and Y. Zhao, "Detection of solar panel defects based on separable convolution and convolutional block attention module," Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, vol. 45, no. 3, pp. 7136-7149, 2023. https://doi.org/10.1080/15567036.2023.2218301

[25] Y. Jie, C. Yong, and Y. Jialin, "Convolutional neural network based on the fusion of image classification and segmentation module for weed detection in alfalfa," Pest Management Science, vol. 80, no. 6, pp. 2751-2760, 2024. https://doi.org/10.1002/ps.7979

[26] X. Xue, C. Zhang, Z. Niu, and X. Wu. "Multi-level attention map network for multimodal sentiment analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 5, pp. 5105-5118, 2023. https://doi.org/10.1145/3517139

[27] X. Zhou, Y. Zhang, Z. Wang, M. Lu, and X. Liu, "MAFN: multi-level attention fusion network for multimodal named entity recognition," Multimedia Tools and Applications,, vol. 83, no. 15, pp. 45047-45058, 2024. https://doi.org/10.1007/s11042-023-17376-5

[28] B. Yu, and S. Zhang, "A novel weight-oriented graph convolutional network for aspect-based sentiment analysis," The Journal of Supercomputing, vol. 79, no. 1, pp. 947-972, 2023. https://doi.org/10.1007/s11227-022-04689-9

[29] J. Zhao, and Y. Li, "Influence of emotional expression in online comments on consumers' perception," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 4, pp. 3343-3352, 2023. https://doi.org/10.1007/s12652-021-03472-7

[30] H. O. Ahmad, and S. U. Umar, "Sentiment analysis of financial textual data using machine learning and deep learning models," Informatica, vol. 47, no. 5, pp. 4673-4678, 2023. https://doi.org/10.31449/inf.v47i5.4673

[31] J. Fernández Herrero, F. Gomezdonoso, and R. Roigvila, "The first steps for adapting an artificial intelligence emotion expression recognition software for emotional management in the educational context," British Journal of Educational Technology, vol. 54, no. 6, pp. 1939-1963, 2023. https://doi.org/10.1111/bjet.13326

[32] R. Ranjan, and A. K. Daniel, "A sentiment ClassificatiCobiat: A sentiment classification model using hybrid Convnet-Dual-lstm with attention mechanismon model using hybrid ConvNet-Dual-LSTM with attention mechanism," Informatica, vol. 47, no. 4, pp. 3911-3924, 2023. https://doi.org/10.31449/inf.v47i4.3911