

Image and Text Aspect Level Multimodal Sentiment Classification Model Using Transformer and Multilayer Attention Interaction

Xiuye Yin, School of Computer Science and Technology, Zhoukou Normal University, China

Liyong Chen, School of Network Engineering, Zhoukou Normal University, China*

ABSTRACT

Many existing image and text sentiment analysis methods only consider the interaction between image and text modalities, while ignoring the inconsistency and correlation of image and text data, to address this issue, an image and text aspect level multimodal sentiment analysis model using transformer and multi-layer attention interaction is proposed. Firstly, ResNet50 is used to extract image features, and RoBERTa-BiLSTM is used to extract text and aspect level features. Then, through the aspect direct interaction mechanism and deep attention interaction mechanism, multi-level fusion of aspect information and graphic information is carried out to remove text and images unrelated to the given aspect. The emotional representations of text data, image data, and aspect type sentiments are concatenated, fused, and fully connected. Finally, the designed sentiment classifier is used to achieve sentiment analysis in terms of images and texts. This effectively has improved the performance of sentiment discrimination in terms of graphics and text.

KEYWORDS

Aspect Level Sentiment Analysis, Deep Attention Interaction Mechanism, Direct Interaction Mechanism of Aspects, ResNet50, RoBERTa-BiLSTM, Transformer

INTRODUCTION

With the rapid development and popularization new media, social networks, and other platforms (Park, 2023), more and more users are inclined to use multimodal forms of data to express their opinions and emotions, with the most used method being the combination of images and text (Bie et al., 2023; Capecchi et al., 2022). Effective emotional analysis of these massive and diverse forms of social media data helps to better understand public emotions and opinion tendencies, thereby providing scientific basis for government and enterprise decision-making (Banik et al., 2023; Wijayanti & Arisal, 2021).

The traditional single modal sentiment analysis method only uses a certain type of information as the analysis object, which cannot meet the needs of multimodal data (Xu et al., 2019). In this situation,

DOI: 10.4018/IJDWM.333854

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

multimodal sentiment analysis has emerged, which will extract and fuse features from multiple modalities of information published by users, thereby more accurately analyzing and predicting their emotions (Huang, Yang et al., 2019; Xiao et al., 2021). However, the existing multimodal sentiment analysis models for image and text fusion still have some shortcomings:

1. Features of different modalities are usually simply concatenated, making it difficult to effectively fuse deep multimodal emotional features.
2. The image information published by social media users may not necessarily be associated with every word in the text. Existing methods do not measure the importance of words in the text based on the specific features of the image, but directly fuse the image and text features, which will directly affect the final emotional classification results.
3. Because aspect-based sentiment analysis belongs to fine-grained sentiment analysis, many existing multimodal sentiment analysis methods lack the ability to solve such problems (Dai et al., 2021).

To address the aforementioned issues, the authors propose an image and text aspect level multimodal sentiment analysis model that based on transformer and multimodal multilayer attention interaction (TF-MMATI). The main contributions are as follows:

1. Using RoBERTa for pretraining and utilizing BiLSTM to fully extract deep semantic information, aiming to better extract features of text modality.
2. To address the issue of ineffective fusion of image and text information, the proposed model aggregates text and image features through Transformer encoder to eliminate the problem of feature differences between text and image modalities.
3. Due to the fact that traditional multimodal image and text sentiment analysis models typically only fuse image and text information, with little consideration given to aspect level, the proposed model utilizes attention interaction mechanisms to weight the features of text modality and image modality at the aspect level respectively, in order to solve the inconsistency problem between image data and text data. At the same time, the researchers solved the correlation problem between text modality and image modality through a deep attention interaction mechanism.

RELATED WORKS

The research closely related to aspect level sentiment analysis of graphic and textual data mainly includes text aspect level sentiment analysis, image sentiment analysis, and graphic and textual sentiment analysis (Alahmary & Al-Dossari, 2023; Li et al., 2022; Mittal & Agrawal, 2022).

Text Level Sentiment Analysis

In the task of sentiment analysis, recurrent neural network and standard attention mechanism are widely used to automatically learn the semantic features of context and aspect words. Hu and Li (2023) proposed a sentiment analysis model based on long short-term memory (LSTM) neural network model, in which they used LSTM-GCN (i.e., graph convolution network) to achieve emotion classification. Shan (2023) proposed a sentiment analysis method based on convolutional neural network (CNN)-bidirectional gated recurrent unit (BiGRU), in which emotional features of different granularity are extracted through CNN, and these emotional features are input into the BiGRU network to get the text emotion type. Mohana et al.'s (2022) study was based on the chaos coyote optimized deep belief network (DBN) for text sentiment analysis. They used a DBN classifier to extract features from the trained data, resulting in an accurate sense of innocence (positive or negative). These methods can provide effective solutions for sentiment analysis based on text information. However, there are

significant bottlenecks in such methods, namely, it is difficult to significantly improve the accuracy of sentiment analysis in the absence of other modal supplements such as images and speech.

Image Sentiment Analysis

There is relatively little research on emotional analysis in the field of images, mainly focusing on two aspects: Facial expression recognition tasks and visual emotion analysis. Arul Vinayakam Rajasimman et al. (2022) devised a novel robust facial expression recognition based on deep learning evolutionary algorithm (i.e., RFER-EADL) model. They used the Dense Net-169 model based on deep CNN, and they adopted chimpanzee optimization algorithm as the super parameter tuning method, using teaching and learning based optimization combined with LSTM model for facial expression recognition and classification. Chen et al. (2020) added a new branch named “texture module” to the traditional CNN. By using this branch and different feature maps to calculate sentiment vectors, they achieved image sentiment differentiation. The above methods can provide effective emotional analysis solutions for facial expression images. However, when applied to ordinary images other than facial expressions, the performance of emotional analysis obtained is not ideal. In addition, in real-world applications such as new media, social networks, and e-commerce, the information provided by users usually includes multiple modalities such as text, images, voice, and video. If sentiment analysis is only based on image modalities, important information from other modalities will be lost, reducing the accuracy of sentiment analysis.

Image and Text Emotional Analysis

In the image and text emotional analysis, the fusion strategy of images and texts is the key for solving the task of emotional analysis of images and texts. According to the fusion strategy of image and text data, it can be divided into feature level fusion, hybrid fusion, and decision level fusion. Feature level fusion is the integration of data from multiple modal sources into a holistic feature vector before emotional scoring. Feature level fusion may generate high-dimensional redundant feature vectors. Decision level fusion occurs after modeling each modality, where the decision results of multiple modal sentiment classifiers are averaged, but the feature interaction between modalities is ignored. Most existing methods for emotional analysis of images and texts adopt a hybrid fusion approach. Liao et al. (2022) proposed an image text interaction graph neural network for image text sentiment analysis, which constructs an image text interaction graph network, combining with the image text information to achieve emotion classification. This method can effectively achieve the analysis of image and text emotions, but it lacks consideration of aspect level information. Huang, Zhang et al. (2019) proposed a new image sentiment analysis model, namely, deep multimodal attention fusion, to use a hybrid fusion framework for sentiment analysis, but ignored the inconsistency of images and texts in the model. Yang et al. (2020) proposed a new multimodal sentiment analysis model based on multiview attention network, which utilizes continuously updated memory networks to achieve sentiment analysis. Khlyzova et al. (2022) contributed a novel annotated English multimodal corpus and developed a model (i.e., TISM) for automatically detecting the relationship between images and texts, as well as the categories of emotional stimuli and emotions. For image text relationships, the information in the text can predict whether images are needed for emotional understanding. Although these two models consider multimodal information, they ignore the aspect level information. Sun and Gao (2021) proposed a multimodal sentiment analysis method that utilizes CNN and attention mechanisms to obtain emotional features and uses bidirectional LSTM for learning to achieve emotion classification of multimodal features. However, these two methods overlook aspect class information. Xu et al. (2018) proposed a new comemory+aspect network, introducing the average value of aspect embedding as input to the text and visual memory network, fully considering image, text, and aspect level information to achieve multimodal aspect like sentiment analysis. However, the emphasis of this method on aspect information needs to be strengthened, and there is a lack of effective feature transformation in the input space, which can lead to issues such as information loss. Yu and

Jiang (2019) proposed a multimodal BERT architecture (i.e., TomBERT), which applies BERT for text representation. At the same time, a set of self-attention layers are stacked at the top to capture multimodal interactions. However, this method does not effectively improve the BERT language model and lack a visible interpretable fusion process. Khan and Fu (2021) designed a dual stream model (i.e., EF-CapTrBERT-DE), which uses an object aware converter to transform images in input space and uses a one-way nonautoregressive text generation method, while extracting multimodal aspect like emotional information using a translation model. However, this method is prone to ignoring inconsistencies between graphic and textual data in given aspects.

Problems and Solutions

Based on the above methods, it is possible to conclude that most existing emotional analysis methods for images and texts usually have the following problems: 1) Ignoring the inconsistency and correlation of image and text data; 2) lack of consideration at the aspect level; 3) deep feature extraction problem. To address these issues, the authors propose an image and text aspect level sentiment analysis model based on TF-MMATL. This model fully integrates aspect information and image and text information from multiple levels, modeling the inconsistency and correlation of image and text data, so that enabling the model to effectively represent the correlation between image and text data and aspect information.

PROPOSED ASPECT LEVEL MULTIMODAL SENTIMENT ANALYSIS MODEL

The sentiment analysis task of image and text aspects can be formally defined as input text content $W = (w_1, w_2, \dots, w_l)$ and image set $U = (u_1, u_2, \dots, u_I)$, predict the emotional labels of a given aspect word $A = (a_1, a_2, \dots, a_n)$ by establishing a model, where l is the length of the text content, I is the number of images, and n is the length of the aspect word. Figure 1 shows the overall structure of the image and text aspect level sentiment analysis model based on transformer and multimodal multilayer attention, including the image and text feature extraction layer, multilevel aspect interaction layer, and image and text feature fusion classification layer. The image and text feature extraction layer is used to extract features of images, texts, and aspects; the multilevel aspect interaction layer is used for multilevel fusion of aspect information and graphic information to remove text and images unrelated to a given aspect, and enhance the emotional representation of the graphic modal data of a given aspect; the image and text feature fusion classification layer is used for concatenation, fusion, and full connection of various emotional representations, achieving sentiment discrimination in terms of image and text.

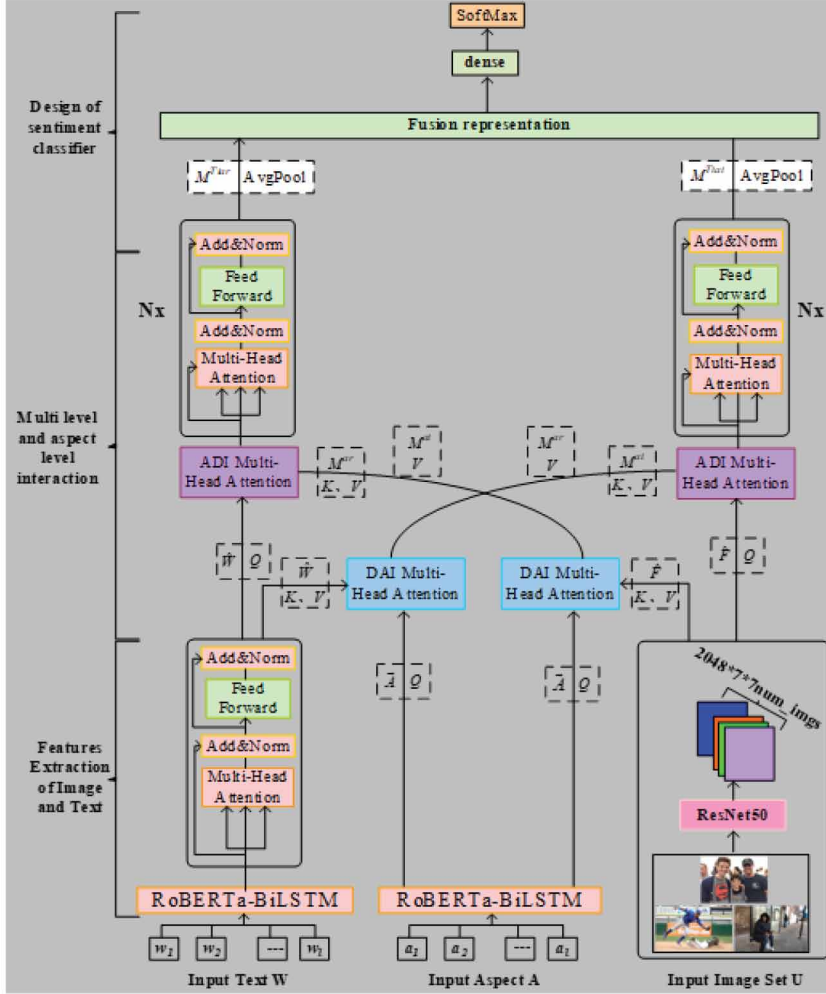
Among them, ResNet50 is used to extract image features, and RoBERTa-BiLSTM is used to extract text features and aspect level features. Through multilayer attention interaction, the emotional attention features of image enhanced text and text enhanced image are fused, and an emotional classifier is designed to output the emotional types of the image and text.

Feature Extraction

Image Feature Extraction

In order to make full use of image information, for the input image set $U = (u_1, u_2, \dots, u_I)$, the proposed model uses CNN ResNet-50 image recognition method to extract its image features (Li & Yan, 2021). 1) Take a specific image $u_i (i = 1, 2, \dots, I)$ from the image set in sequence and readjust it to a uniform pixel size image u'_i ; 2) use pretrained ResNet-50 to obtain image feature vectors \hat{F} . \hat{F} is a third-order tensor, represented as follows:

Figure 1. Overall framework of the proposed TF-MMATl



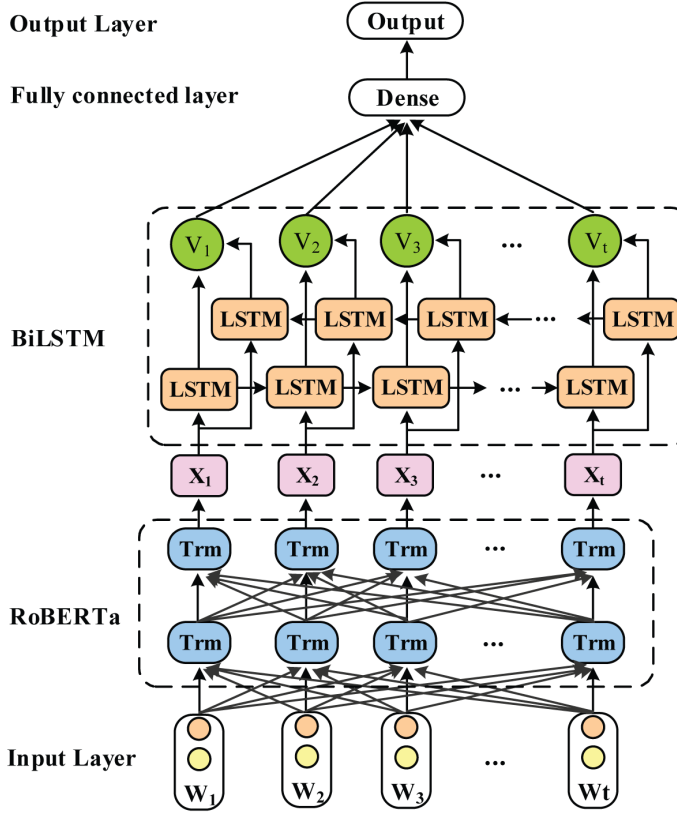
$$\hat{F} = \text{ResNet}(u'_i) \quad (1)$$

Text Feature Extraction

The RoBERTa-BiLSTM model can generate dynamic contextual embedding representations from a large corpus using pretrained model parameters (Figure 2). Given the input text content $W = (w_1, w_2, \dots, w_l)$, l represents the text length, and the embedded representation $\vec{W} = (x_1, x_2, \dots, x_l)$ of the text content is obtained through the RoBERTa-BiLSTM pretraining model, where the vector of each word represents $w_i \in R^k$ and k represents the dimension of the vector.

The image feature level obtained through ResNet-50 is much deeper than the text feature obtained through RoBERTa-BiLSTM model, leading to make the feature representation of the image and text closer. The proposed model utilizes a transform encoder to perform text re-aggregation on the text

Figure 2. Structure of RoBERTa-BiLSTM model



feature \vec{W} obtained from the RoBERTa-BiLSTM model, so that the text features contain deeper semantic information and the reaggregated features are marked as \hat{W} . \hat{W} represents the following:

$$\hat{W} = Te(\vec{W}) \quad (2)$$

where Te represents the transform encoder processing operation.

Aspect Feature Extraction

Similar to text feature embedding representation, the proposed method still utilizes the RoBERTa-BiLSTM model to learn the feature embedding representation of aspect words. Provide the input aspect word $A = (a_1, a_2, \dots, a_n)$, where n represents the length of the aspect word. The RoBERTa-BiLSTM pretraining model is used to obtain the embedding representation $\vec{A} = (b_1, b_2, \dots, b_n)$ of the aspect feature, where the vector of each word representation is $b_i \in R^d$, and d represents the dimension of the vector.

Multilayer Attention Interaction

Due to the inconsistency and correlation between image data and text data, image data should not only provide rich supplementary information, but also avoid introducing aspect independent noise

image and text information (Chen et al., 2022; Zhao et al., 2021). To address these two issues, it is necessary to construct aspect direct interaction mechanisms and aspect deep interaction mechanisms, respectively.

Establishment of Aspect Direct Interaction Mechanisms

To address the inconsistency of image and text modal data, a direct aspect information multihead attention (DAIMA) mechanism is established for direct interaction between aspects. The DAIMA mechanism is used to remove text and images unrelated to a given aspect, while retaining the most relevant parts. The DAIMA mechanism is used again to directly interact with the image representation, text representation, and aspect information of a given aspect. Thus, a visual and textual attention emotional representation of a given aspect is obtained.

For aspect \vec{A} and text content \hat{W} , assuming \vec{A} is a Query and \hat{W} is a Key and Value, the given aspect interacts with the text to obtain the text attention emotion representation M^{AW} of the given aspect as follows:

$$\begin{aligned} M^{AW} &= \text{MultiHead}(\vec{A}, \hat{W}, \hat{W}) \\ &= \text{Concat}(\text{head}_1^A, \dots, \text{head}_h^A) \varpi^{aw} \\ \text{head}_i^A &= \text{Attention}(\vec{A} \varpi_i^{AQ}, \vec{A} \varpi_i^{AK}, \vec{A} \varpi_i^{AV}) \end{aligned} \quad (3)$$

where h is the number of attention heads in multiple heads, and $\varpi^{AW} \in \mathbb{R}^{h \times d_v \times d}$, $\varpi_i^{AQ} \in \mathbb{R}^{d \times d_k}$, $\varpi_i^{AK} \in \mathbb{R}^{d \times d_k}$, $\varpi_i^{AV} \in \mathbb{R}^{d \times d_v}$ are learnable parameters, $d_k = d_v = d / h$.

Similarly, for aspect \vec{A} and image \hat{F} , assuming \vec{A} as Query and \hat{F} as Key and Value, given the interaction between the aspect and the image, the image attention emotion representation M^{AF} limited by the aspect is obtained as follows:

$$\begin{aligned} M^{AF} &= \text{MultiHead}(\vec{A}, \hat{F}, \hat{F}) \\ &= \text{Concat}(\text{head}_1^A, \dots, \text{head}_h^A) \varpi^{af} \\ \text{head}_j^A &= \text{Attention}(\vec{A} \varpi_j^{AQ}, \hat{F} \varpi_j^{AK}, \hat{F} \varpi_j^{AV}) \end{aligned} \quad (4)$$

where h is the number of attention heads in multiple heads, and $\varpi^{AF} \in \mathbb{R}^{h \times d_v \times d}$, $\varpi_j^{AQ} \in \mathbb{R}^{d \times d_k}$, $\varpi_j^{AK} \in \mathbb{R}^{d \times d_k}$, $\varpi_j^{AV} \in \mathbb{R}^{d \times d_v}$ are learnable parameters, $d_k = d_v = d / h$.

Establishment of Deep Interaction Mechanism

Due to the correlation between text data and image data, for the purpose of better facilitating bidirectional interaction between images and texts, and enhancing the emotional information of a given aspect, an aspect deep interaction multihead attention (ADIMA) mechanism is established to fuse aspect information, capture the correlation between text and image information of a given aspect, and further extract accurate semantic and emotional information. On this basis, utilizing the different feature spaces and global feature aggregation capabilities of transformer-encoder, the internal information of modal data is captured (Huang et al., 2020; Lim, 2022).

The proposed model uses text feature embedding representation \hat{W} to query the attention emotion representation M^{AF} of a given aspect of the image modality, so that aspect information deeply participates in the interaction between the image and text. Specifically, the text feature embedding

representation \hat{W} is used as a Query, and the attention emotion representation M^{AF} of a given aspect is used as a Key and Value. For modeling the correlation between images and text, obtaining the emotional representation M^{WAF} of text attention related to the given aspect of the image as follows:

$$\begin{aligned} M^{WAF} &= MultiHead(\hat{W}, M^{AF}, M^{AF}) \\ &= Concat(head_1^W, \dots, head_h^W) \varpi^{waf} \\ head_j^W &= Attention(\hat{W} \varpi_j^{WQ}, M^{AF} \varpi_j^{WK}, M^{AF} \varpi_j^{WV}) \end{aligned} \quad (5)$$

where h is the number of attention heads in multiple heads, and $\varpi^{WAF} \in \mathbb{R}^{h \times d_v \times d}$, $\varpi_j^{WQ} \in \mathbb{R}^{d \times d_k}$, $\varpi_j^{WK} \in \mathbb{R}^{d \times d_k}$, $\varpi_j^{WV} \in \mathbb{R}^{d \times d_v}$ are learnable parameters, $d_k = d_v = d / h$.

The text attention emotion M^{WAF} is related to the given aspect and image, and capturing the single modal internal information of the text through transformer-encoder, namely the text feature information M^{wWAF} , as follows:

$$M^{wWAF} = Te(M^{WAF}) \quad (6)$$

For the obtained text feature information M^{wWAF} , the average pooling is used to obtain the deep interaction text modal feature M_{avg}^{wWAF} , which is represented as follows:

$$M_{avg}^{wWAF} = \sum_{j=1}^n M_j^{wWAF} / n \quad (7)$$

Similarly, obtaining the image modal features of aspect deep interaction, that is, averaging and pooling the obtained image feature information M^{wFAW} , and obtaining the image modal feature M_{avg}^{wFAW} of aspect deep interaction, is represented as follows:

$$M_{avg}^{wFAW} = \sum_{j=1}^n M_j^{wFAW} / n \quad (8)$$

Design of Sentiment Classifier

In order to enriching the emotional semantic information represented by the fusion of image and text modalities, the text modal feature representation M_{avg}^{wWAF} and image modal feature representation M_{avg}^{wFAW} of a given aspect are concatenated and fused, and the image and text information representation S of a given aspect is established as follows:

$$S = Concat(\omega_{wWAF}^T M_{avg}^{wWAF}, \omega_{wFAW}^T M_{avg}^{wFAW}) \quad (9)$$

where ω_{wWAF}^T , ω_{wFAW}^T represents the weights of features M_{avg}^{wWAF} and M_{avg}^{wFAW} . Using S as the fully connected layer input to predict emotional labels for a given aspect.

The proposed method utilizes an emotion classifier to calculate the probability distribution of sentiments and achieve sentiment analysis. Firstly, map the image and text information features S

of the given aspect into the decision space of emotional tendencies. Then, in order to enhance the robustness of the model and improve the classification effect, a bias term was added to adjust the distribution of the given aspect's image and text information feature S (Chawla et al., 2020). Finally, the probability distribution Q is calculated by the activation function softmax (Le et al., 2019). The details are as follows:

$$Q = \text{softmax}(\omega_Q^T S + b_Q) \quad (10)$$

where ω_Q , b_Q is the training weight and bias of the classifier, respectively.

To ensure the consistency of distribution results and the efficiency of optimization, Cross entropy (Xu et al., 2021) is used as the loss function:

$$\text{loss} = -\frac{1}{T} \sum_{i=1}^T \sum_{j=1}^2 Y_j \ln Q_j \quad (11)$$

where Q_j represents the value of the emotional probability distribution Q in the j -th dimension, Y_j represents the value of the encoded real label in the j -th dimension, and T represents the size of each training batch. Thus, by using back propagation, the model can be trained by minimizing the loss function.

EXPERIMENT AND ANALYSIS

Experimental Environment and Datasets

Table 1 shows the experimental environment configuration.

For the purpose of evaluating the proposed model, the authors used two target oriented multimodal sentiment classification benchmark datasets Twitter-15 and Twitter-17 (Khan & Fu, 2021). Twitter 15 and 17 consist of multimodal tweets, with each multimodal Twitter consisting of text, images posted next to the tweet, the target in the tweet, and the emotions of each target. Each target is assigned a label in the set {positive, neutral, negative}, which is a standard multiclass classification problem (Ye et al., 2021). Download links for the Twitter 15 and 17 datasets are at <https://dev.twitter.com>.

Parameter Settings

In order to achieve better classification results on the dataset, the authors defined some parameter settings in the experiment (Table 2). For the image, they adjusted its size parameter to 224*224, and used the pretrained ResNet-50 to obtain a 7*7*2048-dimensional visual feature vector. For the proposed model, Batch Size represented the number of training samples in each batch, the two

Table 1. Experimental environment settings

Experimental environment	Configuration
Programming language	Python 3.8
Deep learning framework	Pytorch 1.10
Operating system	Ubuntu1.04
GPU	NVIDIA Tesla A100 80GB

Table 2. Parameter settings of the model

Modal	Parameter	Value
Image	Image size	224*224
	ResNet output	7*7*2048
Text	Maximum length of text	50
	Number 1 of BiLSTM hidden units	128
	Number 2 of BiLSTM hidden units	128
	Batch size	32
Others	Dropout	0.5
	Learning rate	$1*10^{-3}$
	Number of attention heads	2
	Epoch	50

layers of the BiLSTM network layer set the same number of hidden layer units for LSTM, and Epoch represented the number of training iterations. The authors used the dropout method in the experiment to alleviate overfitting. Considering the limited computing power of the computing platform used in this experiment, they set the Batch Size value to 32. They set the number of hidden units in the BiLSTM network layer to 64, 128, and 256 values during the experiment, and the classification effect was best when its value was 128. Meanwhile, the researchers set the number of attention heads to 2, and the optimizer used Adam. The authors used the early stop method during the training process for obtaining a model with strong generalization ability and preventing overfitting, which means that, when the accuracy of the validation set and F1 value remain unchanged for 10 consecutive rounds, the training process will stop.

Evaluation Index

For the three categories of positive (P), negative (N), and neutral (M) emotions, due to the imbalance of data sample categories, the authors use the idea of weighted average to calculate four indicators: Precision, recall, F1 score, and accuracy. The specific calculation is as follows:

$$\begin{aligned}
 Accuracy &= \frac{TP + TN + TM}{TP + TN + FP + FN + TM + FM} \\
 Precision &= \sum_{g \in \{P, N, M\}} \frac{TP_g}{TP_g + FP_g} \mu_g \\
 Recall &= \sum_{g \in \{P, N, M\}} \frac{TP_g}{TP_g + FN_g} \mu_g \\
 F1 &= \sum_{g \in \{P, N, M\}} F1_g \times \mu_g
 \end{aligned} \tag{12}$$

where T and F represent the situations where the predicted value is equal to the label value and the predicted value is not equal to the label value, respectively. μ_g is the proportion of category samples. The calculations of precision, recall, and F1 score are combined with the one-vs.-all strategy. Assuming that the positive category (P) is considered positive (i.e., $g=P$), the results of both the negative (N) and neutral (M) categories are considered negative.

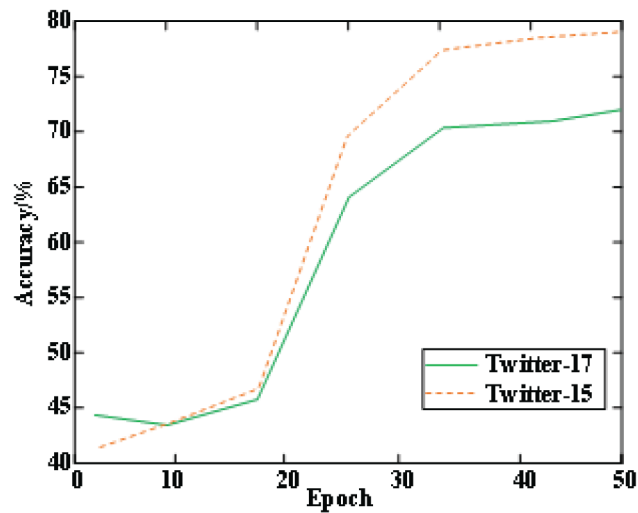
Model Training

Convergence Analysis of the Model

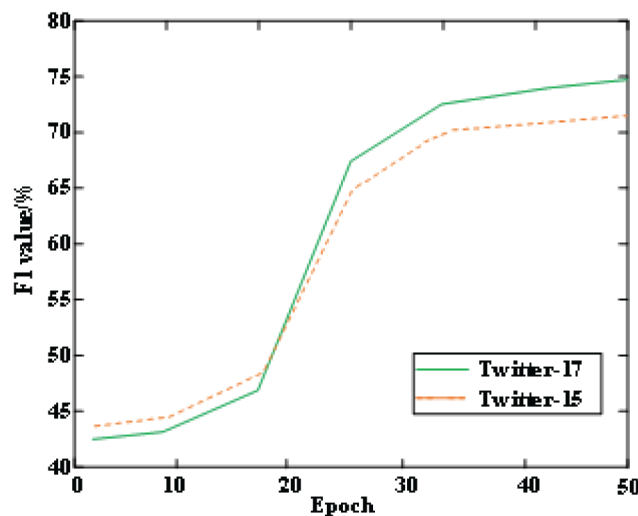
Figure 3 shows the accuracy and F1 values of the proposed model on two datasets, Twitter 15 and 17, as a function of the number of iterations.

Figure 3 evidences that, before the 20th iteration, the accuracy and F1 value of the proposed model showed a slow upward trend overall because the model had just started training, and the weight parameters were in the initialization stage. It is necessary to increase the number of iterations to train parameters that can improve the model's classification performance. After 20 iterations, the accuracy

Figure 3. Convergence curve of the proposed model



(a)-Accuracy



(b)-F1

and F1 value of the proposed model generally showed a rapid upward trend on both datasets. The reason is that, after iterative training, the model has already trained some effective parameters that can improve the model's classification performance, which leads to accelerating the iteration speed. After the 35th iteration, the proposed model was generally in a stable state in terms of accuracy and F1 value, as the model needs to reach a fitting state and the results do not change significantly. Finally, on the Twitter-15 and Twitter-17, the accuracy and F1 values of the proposed model were tended to stabilize, and the model reached convergence.

Loss Training of the Model

Due to the design of an emotion classifier for the proposed model, the loss value at optimal convergence differs significantly from other models. Therefore, the authors further analyzed the convergence of the proposed model on the Twitter 15 and 17 through the normalized loss value convergence curve of 50 Epoch iterations (Figure 4).

Figure 4 shows that, after 35 Epochs, the proposed model quickly converges, with a loss value approaching 0.1, and the training effect is ideal. Meanwhile, on the Twitter 15 and 17, the loss values of the proposed model have similar trends, indicating that the classification results of the proposed model have a certain degree of stability when faced with differential data.

Model Parameter Analysis

Parameter Analysis of the Dropout

Aiming to further improve the analysis accuracy of the proposed model, the dropout method is used in the experiment to alleviate overfitting, but different values of dropout will directly affect the output results of the model. For the purpose of setting reasonable dropout values, the authors conducted multiple experiments on two datasets, Twitter 15 and 17. Figure 5 shows the indicator values obtained by the proposed model.

Figure 4. Convergence curve of normalized loss value

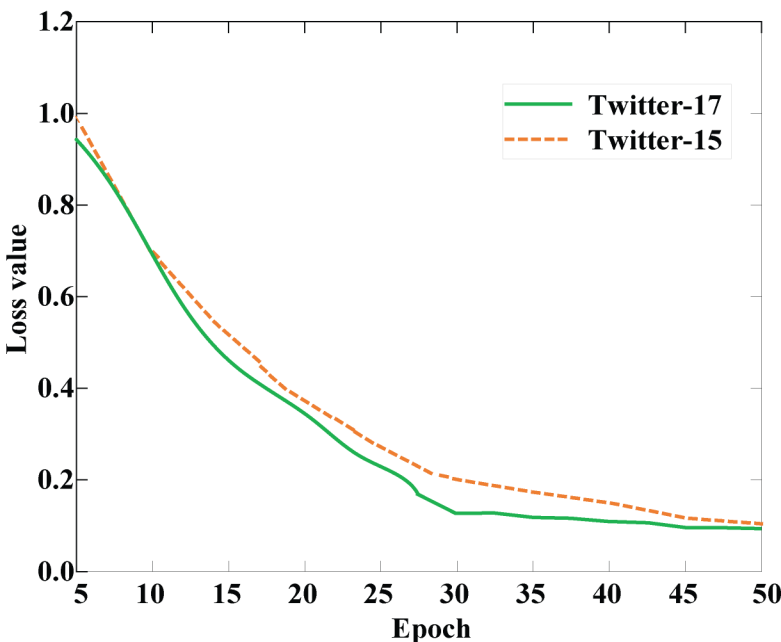
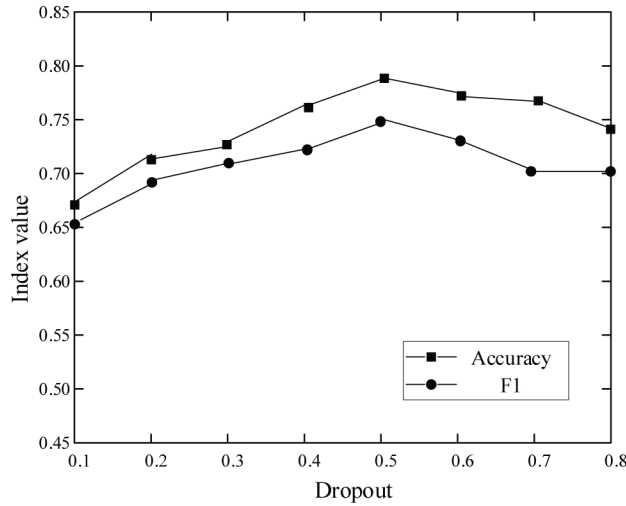
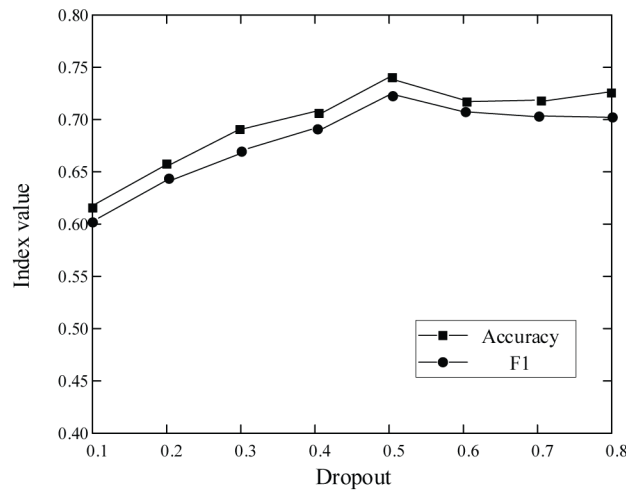


Figure 5. Influence curve of dropout values on the model



(a) Twitter-15



(b) Twitter-17

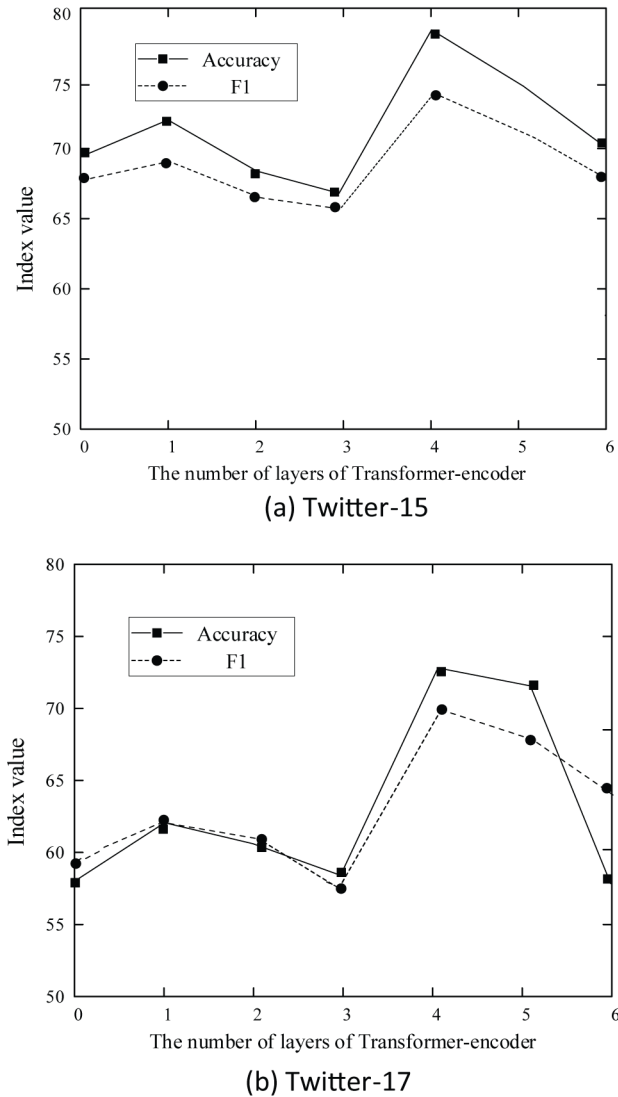
As Figure 5 illustrates, when the dropout value is 0.5, the proposed model achieves the highest accuracy and F1 value on both datasets, improving the model's generalization ability. Therefore, the dropout parameter takes a value of 0.5.

Layer Number Effect of Transformer-Encoder

In the proposed model, the internal information of a single mode is captured through a transformer-encoder, and the number of layers of the transformer-encoder has a significant impact on the results of sentiment analysis. Therefore, the authors conducted experiments on the Twitter 15 and 17 with the number of layers of transformer-encoder. Figure 6 shows the accuracy and F1 values obtained by the proposed model.

Figure 6 shows that, when the number of transformer-encoder layers is set to 4, the proposed model achieves the best aggregation ability. Taking the Twitter 17 as an example, its accuracy and

Figure 6. Influence curve of the transformer-encoder



F1 value are both close to 77%. This indicates that the sentiment analysis model incorporating a four-layer transformer encoder can fully capture the internal information of the modality.

Comparison and Analysis With Other Advanced Models

The authors used the following model for comparative experiments:

1. **AEAT-LSTM (Hu & Li, 2023):** This model adds aspect embedding in each word embedding and combines the word hiding state with aspect embedding in the attention layer to focus on keywords associated with a given aspect.
2. **DBN (Mohana et al., 2022):** A DBN classifier is used to extract features from the trained data, resulting in an accurate sense of innocence (positive or negative).

3. **CNN-BiGRU (Shan, 2023):** This model extracts emotional features of different granularity through CNN, which will be input into BiGRU network for text emotion type analysis.
4. **TISM (Sun & Gao, 2021):** This model automatically detects the relationship between images and text, as well as the categories of emotional stimuli and emotions.
5. **Comemory+aspect (Xu et al., 2018):** In addition to modeling interactive attention between text and visual memory, this model also introduces the average value of aspect embedding as input to the text and visual memory network.
6. **TomBERT (Yu & Jiang, 2019):** This model includes a multimodal BERT architecture, which uses BERT to obtain text Word embedding, and then cross modal attention to focus on the target word image, to obtain target image attention embedding, and to form superimposed multilayer self-attention.
7. **EF-CapTrBERT-DE (Khan & Fu, 2021):** This dual stream model uses an object aware converter to transform images in input space and uses a one-way nonauto regressive text generation method, while extracting multimodal aspect class information using a translation model.

The evaluation indicators the authors used in the experiment were accuracy and F1 value. The researchers compared the above models in the Twitter 15 and 17; Table 3 shows the results.

The experimental results in Table 3 evidence the following:

1. The overall performance of the multimodal sentiment analysis model is superior to that of the single modal model. This is due to the fact that the methods based solely on text modality can only extract emotional information from text, but they cannot extract the emotional information contained in the image. For DBN, ATAE-LSTM, and CNN-BiGRU single mode models, the analysis accuracy is less than 66%. At the same time, compared to DBN on the Twitter-15 dataset, ATAE-LSTM performs better with an accuracy improvement of 1.28%. The reason is that ATAE-LSTM utilizes multiple attention layers to obtain more comprehensive text information. Compared to the single modal model, the proposed model and other multimodal models can fuse aspect information and graphic information at multiple levels by combining text and image data, thereby retaining more and more important aspect emotional information, and significantly improving recognition performance. Taking the Twitter-17 dataset as an example, the accuracy of the proposed TF-MMATI model is improved by 11.85%, compared to the CNN BiGRU model.

Table 3. Comparison of evaluation index values of different models

Model	Twitter-15		Twitter-17	
	Accuracy/%	F1/%	Accuracy/%	F1/%
Text only				
DBN	61.85	61.47	53.59	50.38
AEAT-LSTM	63.13	62.68	56.01	55.24
CNN-BiGRU	66.44	65.93	60.82	60.33
Text + image				
TISM	68.77	68.23	63.67	63.09
Comemory+aspect	74.12	69.85	68.53	66.01
TomBERT	77.15	71.75	70.34	68.03
EF-CapTrBERT-DE	77.92	73.90	72.30	70.20
TF-MMATI (the authors')	78.66	74.53	72.67	70.82

2. The analysis results of the proposed TF-MMATI are superior to other contrastive multimodal fusion models of image and text. Although the TISM model combines text and image modalities, it fails to take into account aspect-based information well, resulting in lower performance in aspect-based sentiment analysis, compared to other multimodal models. The comemory+aspect model combines text, visual, and aspect information, which can effectively analyze emotions. However, due to the lack of effective feature transformation in the input space, this model is prone to losing the aspect information inherent in the image itself. Compared to comemory+aspect, the TomBERT model and EF-CapTrBERT-DE can significantly achieve better results, but both are slightly lower than the proposed TF-MMATI model. TomBERT does not make substantial improvements to the BERT language model, and, more importantly, the model lacks an interpretable feature fusion process, where the correlation between text modality, visual modality, and given aspect words still needs to be strengthened. The EF-CapTrBERT-DE model to some extent ignores the inconsistency between given aspect graphic data. The proposed TF-MMATI fully utilizes the inconsistency between text modalities, image modalities, and the correlation with aspect words in the given context, and it interacts aspect information with graphic and textual information at multiple levels. Compared to other multimodal fusion models for comparison, the proposed TF-MMATI utilizes attention interaction mechanisms to weight the features of text modality and image modality at the aspect level respectively, which can more fully obtain aspect information in images and texts, and further capture internal information of modalities using transformer-encoder. The F1 values on the Twitter-15 and Twitter-17 datasets can reach as high as 74.53% and 70.82%, respectively.

Ablation Experiment

Aiming to evaluating the rationality and effectiveness of the proposed module, the authors conducted ablation experiments on two datasets, namely, Twitter 15 and 17 (Table 4). The ablation experimental design is as follows:

- **Without (w/o) ResNet50:** It represents removing the image part from the model.
- **w/o RoBERTa-BiLSTM:** It represents removing the text part from the model.
- **w/o DAIMA:** It denotes the removal of multimodal attention emotional representations of a given aspect obtained through the DAIMA mechanism.
- **w/o ADIMA:** It represents the removal of the text modal feature representation and image modal feature representation obtained through the ADIMA mechanism.
- **w/o Transformer-Encoder:** It refers to removing the part of single mode information captured through transformer encoder.

In Table 4, the performance of the proposed models on the Twitter 15 and 17 is superior, with accuracy rates of 77.86% and 72.29%, respectively. The experimental results of w/o ResNet50 and w/o RoBERTa-BiLSTM show that using only single mode image and text data cannot effectively solve the inconsistency between image and text data, leading to a decrease in the accuracy of aspect-based sentiment analysis, especially the lack of text data processing. Taking the Twitter-15 dataset as an example, the accuracy and F1 are only 38.32% and 30.36%, respectively. The w/o DAIMA model loses the capture of inconsistencies between graphic and textual data during the interaction process, while also lacking direct interaction of aspect information, leading to the introduction of noise unrelated to aspect information. The w/o ADIMA model lacks the interaction between deep aspect information and textual information, resulting in insufficient interaction between textual information in a given aspect. Both of these models have reduced the accuracy of aspect-based sentiment analysis. Taking F1 as an example, the results on the Twitter 17 dataset are 74.18% and 70.59%, respectively. Due to the fact that not using a transformer-encoder will reduce the aggregation ability between features and

Table 4. Results of ablation experiment




Data set		Twitter-15	Twitter-17
TF-MMATI (the authors')	Accuracy/%	78.66	72.67
	F1/%	74.53	70.82
w/o ResNet50	Accuracy/%	65.04	59.95
	F1/%	64.81	59.47
w/o RoBERTa-BiLSTM	Accuracy/%	38.32	33.08
	F1/%	30.36	32.74
w/o DAIMA	Accuracy/%	76.57	70.52
	F1/%	72.99	68.87
w/o ADIMA	Accuracy/%	77.65	70.93
	F1/%	74.18	70.59
w/o Transformer-encoder	Accuracy/%	78.03	71.96
	F1/%	73.89	70.52

cannot capture the internal information of a single mode, the analysis effect will be reduced, but the overall impact on the proposed TF-MMATI is minimal.

Sample Analysis

For the purpose of further demonstrating the effectiveness of the proposed model, the authors selected three sample instances from Twitter for sentiment analysis probability calculations (Table 5). The emotional analysis results are encoded as [positive, neutral, negative], examples belonging to this

Table 5. Sample analysis results

Sample	1	2	3
Image			
Text	A woman is sitting on a sidewalk.	A baseball player sliding into a base.	A group of young men standing next to each other.
Specific aspect	Daily life	Game	Socialize
Label	[0,0,1]	[0,1,0]	[1,0,0]
TomBERT	[0.0,0.163,0.837]	[0.042,0.72,0.196]	[0.926,0.074,0]
EF-CapTrBERT-DE	[0.005,0.211,0.784]	[0.081,0.662,0.257]	[0.905,0.095,0]
TF-MMATI (the authors')	[0,0.051,0.949]	[0.036,0.879,0.085]	[0.983,0.017,0]

type are labeled as “1,” and examples that do not belong to this type are labeled as “0.” Table 5 shows the sample analysis results.

The performance of sample 3 in Table 5 evidences that, for samples with strong “positive” emotions in social interaction, all models can accurately judge, with probability values exceeding 0.900. For the “negative” samples in sample 1’s daily life, all three models show some judgment bias, especially TomBERT, whose probability of judging negative emotions is only 0.784. Although TomBERT combines image, text, and aspect level information, it has not effectively improved the BERT language model and lacks an interpretable fusion process that can be displayed. The extraction of emotional features from specific aspects such as daily life is not comprehensive enough, resulting in low accuracy. Due to the requirement of fairness and impartiality in the competition, objective response should not be carried out with personal emotions. For the “neutral” samples in the sample 2 competition, all three models show significant misjudgments, but the proposed TF-MMATI has the highest accuracy calculation probability of 0.879. This is because the TF-MMATI used a multilayer attention interaction mechanism to fuse image and text information, and the RoBERTa-BiLSTM model is used to extract aspect level features, further ensuring the accuracy of the analysis. The performance variation of the proposed TF-MMATI on different samples is relatively small, which proves that it can more accurately and effectively combine emotional information in images and texts, and it has stronger stability.

CONCLUSION

An image and text aspect level sentiment analysis model based on transformer and multimodal multilayer attention interaction is proposed. By using the proposed TF-MMATI, the inconsistency and correlation between the image and text data of a given aspect are modeled separately, enhancing the representation of the image and text modal data of the given aspect. The accuracy rates obtained on the Twitter-15 and Twitter-17 datasets are 78.66% and 72.67%, respectively, and the F1 values are 74.53% and 70.82%, respectively.

However, the proposed TF-MMATI also has certain limitations. For example, when the dataset has issues such as mixed language, short length or missing word errors, the proposed model is difficult to achieve satisfactory results. In future work, the authors will explore new feature extraction methods to ensure accurate extraction of emotional information on multimodal datasets of different sizes and types. In addition, the authors will consider information such as social relationships for more effective emotional analysis.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China (U1404622), the Key Scientific and Technological Project of Henan Province (232102210075, 212102210098, 212102210400), and the Key Research Projects of Henan Provincial Department of Education (23A520045, 22A520051).

REFERENCES

- Alahmary, R., & Al-Dossari, H. (2023). A semiautomatic annotation approach for sentiment analysis. *Journal of Information Science*, 49(2), 398–410. doi:10.1177/01655515211006594
- Arul Vinayakam Rajasimman, M., Manoharan, R. K., Subramani, N., Aridoss, M., & Galety, M. G. (2022). Robust facial expression recognition using an evolutionary algorithm with a deep learning model. *Applied Sciences (Basel, Switzerland)*, 13(1), 468. doi:10.3390/app13010468
- Banik, D., Satapathy, S. C., & Agarwal, M. (2023). Advanced weighted hybridized approach for recommendation system. *International Journal of Web Information Systems*, 19(1), 1–18. doi:10.1108/IJWIS-01-2022-0006
- Bie, Y., Yang, Y., & Zhang, Y. (2023). Fusing syntactic structure information and lexical semantic information for end-to-end aspect-based sentiment analysis. *Tsinghua Science and Technology*, 28(2), 230–243. doi:10.26599/TST.2021.9010095
- Capecchi, I., Barbierato, E., & Bernetti, I. (2022). Analyzing TripAdvisor reviews of wine tours: An approach based on text mining and sentiment analysis. *International Journal of Wine Business Research*, 34(2), 212–236. doi:10.1108/IJWBR-04-2021-0025
- Chawla, P., Hazarika, S., & Shen, H. W. (2020). Token-wise sentiment decomposition for ConvNet: Visualizing a sentiment classifier. *Visual Informatics*, 4(2), 132–141. doi:10.1016/j.visinf.2020.04.006
- Chen, H., Pei, J. H., & Zhao, Y. (2022). Pedestrian sequence attribute recognition method with multi-feature fusion combined with temporal attention mechanism. *Journal of Signal Processing*, 38(1), 64–73.
- Chen, J., Mao, Q. R., & Xue, L. Y. (2020). Visual sentiment analysis with active learning. *IEEE Access : Practical Innovations, Open Solutions*, 8, 185899–185908. doi:10.1109/ACCESS.2020.3024948
- Dai, Z., Liu, Y., Di, S., & Fan, Q. (2021). Aspect-level sentiment analysis merged with knowledge graph and graph convolutional neural network. *Journal of Physics: Conference Series*, 2083(2), 042044. doi:10.1088/1742-6596/2083/4/042044
- Hu, J. P., & Li, Y. G. (2023). Research on the application of LSTM neural network model in text sentiment analysis and sentiment word extraction. *Advances in Computer, Signals, and Systems*, 7(2), 63–70.
- Huang, F. R., Zhang, X. M., Zhao, Z. H., Xu, J., & Li, Z. (2019). Image-text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167, 26–37. doi:10.1016/j.knosys.2019.01.019
- Huang, W., Mao, Y., Yang, Z., Zhu, L., & Long, J. (2020). Relation classification via knowledge graph enhanced transformer encoder. *Knowledge-Based Systems*, 206, 106321. doi:10.1016/j.knosys.2020.106321
- Huang, Y., Yang, J., Liu, S., & Pan, J. (2019). Combining Facial Expressions and Electroencephalography to Enhance Emotion Recognition. *Future Internet*, 11(5), 105. doi:10.3390/fi11050105
- Khan, Z., & Fu, Y. (2021). Exploiting BERT for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, (pp. 3034–3042). Association for Computing Machinery. doi:10.1145/3474085.3475692
- Khlyzova, A., Silberer, C., & Klinger, R. (2022). On the complementarity of images and text for the expression of emotions in social media. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, (pp. 1–15). Association for Computational Linguistics. doi:10.18653/v1/2022.wassa-1.1
- Le, C. C., Prasad, P. W., Alsadoon, A., Pham, L., & Elchouemi, A. (2019). Text classification: Naive Bayes classifier with sentiment lexicon. *IAENG International Journal of Computer Science*, 46(2PT.141-263), 141–148.
- Li, S., & Yan, W. (2021). Resnet 50 based method for cholangiocarcinoma identification from microscopic hyperspectral pathology images. *Journal of Physics: Conference Series*, 1880(1), 012019–012026. doi:10.1088/1742-6596/1880/1/012019
- Liao, W., Zeng, B., Liu, J., Wei, P., & Fang, J. (2022). Image-text interaction graph neural network for image-text sentiment analysis. *Applied Intelligence*, 52(10), 11184–11198. doi:10.1007/s10489-021-02936-9

- Lim, H. (2022). BERTOEIC: Solving TOEIC problems using simple and efficient data augmentation techniques with pretrained transformer encoders. *Applied Sciences (Basel, Switzerland)*, 12(13), 6686–6693. doi:10.3390/app12136686
- Mittal, D., & Agrawal, S. R. (2022). Determining banking service attributes from online reviews: Text mining and sentiment analysis. *International Journal of Bank Marketing*, 15(3), 40–51. doi:10.1108/IJBM-08-2021-0380
- Mohana, R. S., Rajathi, K., Kousalya, K., & Yuvaraja, T. (2022). Text sentiment analysis on e-shopping product reviews using chaotic coyote optimized deep belief network approach. *Concurrency and Computation*, 34(19), e7039. doi:10.1002/cpe.7039
- Park, C.-S. (2023). Efficient keyword search on graph data for finding diverse and relevant answers. *International Journal of Web Information Systems*, 19(1), 19–41. doi:10.1108/IJWIS-09-2022-0157
- Shan, Y. C. (2023). Social network text sentiment analysis method based on CNN-BiGRU in big data environment. *Mobile Information Systems*, 2023, 8920094. doi:10.1155/2023/8920094
- Sun, Y., & Gao, J. (2021). Fine-grained multimodal sentiment analysis based on gating and attention mechanism. *Electronics Science Technology and Application*, 7(4), 123. doi:10.18686/esta.v7i4.166
- Wijayanti, R., & Arisal, A. (2021). Automatic Indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1), 1–16. doi:10.1145/3425632
- Wu, D. C., Zhong, S., Qiu, R. T., & Wu, J. (2021). Are customer reviews just reviews? Hotel forecasting using sentiment analysis. *Tourism Economics*, 28(3), 795–816. doi:10.1177/13548166211049865
- Xiao, X., Yang, J., & Ning, X. (2021). Research on multimodal emotion analysis algorithm based on deep learning. *Journal of Physics: Conference Series*, 1802(3), 032054,10.
- Xu, J., Huang, F., Zhang, X., Wang, S., Li, C., Li, Z., & He, Y. (2019). Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowledge-Based Systems*, 178, 61–73. doi:10.1016/j.knsys.2019.04.018
- Xu, N., Mao, W., & Chen, G. D. (2018). A co-memory network for multimodal sentiment analysis. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 929–932). Association for Computing Machinery. doi:10.1145/3209978.3210093
- Xu, X., Li, Y., & Yuan, C. (2021). Conditional image generation with one-vs-all classifier. *Neurocomputing*, 434(28), 261–267. doi:10.1016/j.neucom.2020.12.091
- Yang, X., Feng, S., Wang, D., & Zhang, Y. (2020). Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23, 4014–4026. doi:10.1109/TMM.2020.3035277
- Ye, S., Ho, K., & Zerbe, A. (2021). The effects of social media usage on loneliness and well-being: Analysing friendship connections of Facebook, Twitter, and Instagram. *Information Discovery and Delivery*, 49(2), 136–150. doi:10.1108/IDD-08-2020-0091
- Yu, J. F., & Jiang, J. (2019). Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, (pp. 5408–5414). IEEE. doi:10.24963/ijcai.2019/751
- Zhao, Q., Liu, J., Li, Y., & Zhang, H. (2021). Semantic segmentation with attention mechanism for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13. doi:10.1109/TGRS.2020.3042202

Xiuye Yin is a lecturer at Zhoukou Normal University. She received her master's degree from Liaoning University of Science and Technology, China. Her research focus is artificial intelligence.

Liyong Chen is an associate professor at Zhoukou Normal University. He received his master's degree from the School of Computer Science and Technology, Faculty of Electronic Information, Liaoning University of Science and Technology, China, in 2010. His research focus is artificial intelligence.