



**International Journal of Information and Communication Technology**

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

---

**Emotion recognition of consumer comments based on graphic fusion**

Li Lin, Minglan Yuan, Shangzhen Pang, Rongping Wang

**Article History:**

Received:	24 October 2024
Last revised:	21 November 2024
Accepted:	21 November 2024
Published online:	02 January 2025

---

## Emotion recognition of consumer comments based on graphic fusion

---

Li Lin\*

Finance and Economics Department,  
Nanchong Vocational and Technical College,  
Nanchong, Sichuan 637000, China  
Email: linli@nczy.edu.cn

and

Graduate School of Business (GSB),  
SEGi University,  
Kota Damansara 47810, Malaysia

\*Corresponding author

Minglan Yuan

School of Economic Management and Public Affairs,  
Chongqing Vocational College of Architecture and Engineering,  
Chongqing 400072, China  
Email: yuan\_ml0926@126.com

Shangzhen Pang

School of Physics and Electronic Engineering,  
Sichuan University of Science and Engineering,  
Zigong, Sichuan 643000, China  
Email: psz10@126.com

Rongping Wang

School of Finance and Trade Management,  
Chengdu Vocational and Technical College of Industry and Trade,  
Chengdu, Sichuan 611731, China  
Email: 870508647@qq.com

**Abstract:** Multimodal sentiment classification is the use of image and text data to mine the potential emotions in consumer reviews. The picture scenes of the comments are complex, and there is a lot of information affecting the emotional judgement. To solve this problem, we propose a multimodal sentiment analysis model based on image translation to eliminate the redundant features of images and improve the efficiency of the model. Enhance text data to generate diverse data. Interactive learning realises the deep fusion of image and text. Leverage transformer to convert images into text by inputting them into space and output

accurate consumer sentiment information in the images. Construct auxiliary sentences through translation, increase the amount of available text, and output classification results. Experiments were carried out on multi-ZOL, Twitter-2015 and 2017 datasets, and the accuracy rate reached 71.94%, 79% and 88.35%, respectively, which is far superior to other advanced methods.

**Keywords:** emotion analysis; picture translation; transformer; consumer reviews; text enhancement.

**Reference** to this paper should be made as follows: Lin, L., Yuan, M., Pang, S. and Wang, R. (2024) 'Emotion recognition of consumer comments based on graphic fusion', *Int. J. Information and Communication Technology*, Vol. 25, No. 12, pp.46–65.

**Biographical notes:** Li Lin, Associate Professor, Dr. in Reading of Philosophy (Management) from SEGi University, Master of Project Management with Merit graduated from University of Greenwich in 2007. She works at Nanchong Vocational and Technical College. Her research interests include e-commerce, new media marketing, digital marketing, business decision analysis.

Minglan Yuan, Professor, Master of Software Engineering, Graduating from University of Electronic Science and Technology of China in Sichuan Province in 2012. She works at Chongqing Jianzhu College. Her research interests include software technology and network.

Shangzhen Pang, Lecturer, Dr. in Reading of Philosophy (Engineering) from SEGi University, Master, graduated from Chongqing University in 2004. She works at Sichuan University of Science and Engineering. Her research interests include Signal and information processing, electronic circuit design.

Rongping Wang, Professor, Master, graduated from Webster University in August 2022. Currently, he is working at Chengdu Industry and Trade College. His research fields include vocational education, economic management and e-commerce.

---

## 1 Introduction

Consumers around the world can use virtual networks as a medium to share their reviews, insights and opinions about products (Bening et al., 2023; Mubeen et al., 2022). Different from traditional paper media, the forms of content expression of online social media are more complex and diversified. The content of consumer reviews is no longer limited to text description, but provides more vivid and three-dimensional information content with the help of digital media technologies such as image, voice and video (Roohani and Vincheh, 2023; Braghieri et al., 2022). Most of the early sentiment analysis work focused on textual data to build textual feature representations. However, a single text sentiment analysis technique cannot adapt to the complex multimodal environment in online social media. In the product review scenario, users attach multiple images to their reviews.

Images are often closely related to text content and can provide important additional information in sentiment analysis tasks, helping to locate the position of emotion words and strengthening the expression of emotions. Both pictures and texts have emotional characteristics, which can explain the complete emotional state from different perspectives (Bakhit et al., 2024; Omuya et al., 2023). Although multi-modal data can provide richer feature information, the heterogeneity of multi-modal data brings new problems and challenges to sentiment analysis tasks. Different modes need to be processed by corresponding methods to extract their feature representations. Relevant studies have proved that in multi-modal sentiment analysis tasks, shared information between modes and unique information within modes are essential information sources (Ying et al., 2023; Bhatt et al., 2021). The key of feature fusion is to eliminate modal noise and redundant information and to discover the interactive information between modes. According to different application scenarios, a suitable model needs to be designed to implement the task of sentiment analysis.

The multimodal model uses high-level features of images and text for sentiment analysis and global information alignment between different modes (Hou et al., 2023). The fusion effect between different modes of the model is poor, resulting in poor recognition effect. Cross-attention and self-attention modules are introduced into the sentiment analysis model, and convolution layers are used to extract local and global features, effectively improving the recognition accuracy (Zhang, 2022). However, the convolution layer and cyclic layer in the model have the problem of gradient disappearing or explosion, which is a challenge for the model. Zhang and Zhang (2022) used improved CNN for text feature extraction, improved dense block method for image feature extraction, and finally used attention mechanism fusion mode for emotion analysis. Compared with traditional sentiment analysis methods, the model's performance has been significantly improved. However, the improved dense block method increases the training time and decreases the generalisation ability. Support vector machines, deep neural networks, and gradient boosted decision trees are also commonly used for sentiment analysis. The performance of these three preference learning models is significantly better than that of traditional baselines, and the decision tree model with gradient enhancement has the best performance (Lei and Cao, 2023). However, the effect of gradient enhanced decision trees in processing large-scale data is not satisfactory.

To sum up, this paper mainly carries out a series of research and innovation from the aspects of feature representation, information fusion, model design and so on. Based on image and text modal information, a multi-modal sentiment analysis model based on spatial translation is implemented. The main innovation is as follows:

A new structure for image and text emotion recognition is proposed. The multimodal sentiment analysis model (MSST) of consumer reviews based on image translation is designed into three key modules, namely, data processing module, image translation module and classification module. The architecture performs translation in the input space and fuses auxiliary sentences using a large pre-trained language model. Transformer is used to generate text. The trained model is used to convert the image input space into text and construct auxiliary sentences. BERT language model is used for multimodal sentiment analysis. In our extensive experiments, three public datasets are used: Twitter-2015 (Dalmia et al., 2015), Twitter-2017 (Yang et al., 2022) and Multi-ZOL (Liu et al., 2022). Experimental results show that the method proposed in this

paper is significantly better than other related studies. The ablation analysis showed that each component of the model helped to efficiently classify emotions.

The rest of the paper is structured as follows: Section 2 provides an overview of the background work on MSST. Section 3 introduces the new MSST model approach. Section 4 provides a detailed analysis of related experiments and their results. Finally, Section 5 briefly summarises the main points of this paper.

## **2 Related work**

With the progress of science and technology (Jaiswal and Arun, 2021), the functions and services provided by social media become more and more abundant, and the content posted by users also shows a trend of diversification. Social media platforms focusing on photo sharing began to emerge (Chauhan et al., 2021; Zhang et al., 2022). Facebook, Instagram and Tumblr, for example, have become increasingly popular among young people in recent years. Traditional blog platforms and review sites have also followed suit, upgrading traditional services and adding support for photo uploading and sharing, such as Sina Weibo's photo blog and Dianping's photo comments.

As a sub-problem of multimodal sentiment analysis, feature extraction is the primary task of graphic sentiment analysis. At the same time, the cross-modal graphic fusion is the integration of the original information, and also the key to construct the joint feature representation and realise the emotion analysis.

Image emotion feature extraction method: most of the early researches were based on the theoretical knowledge of visual psychology and colour psychology, and explored the relationship between image colour, texture, shape contour and other potential visual features and emotional states (Tuncer et al., 2022). Kode and Barkana, (2023) evaluated three image feature extraction methods of convolutional neural network, transfer learning VGG16 and knowledge-based system and their performance in medical diagnosis, among which knowledge-based feature extraction has the best effect. An unsupervised learning framework for spectral motion features is proposed. Learn to perceive spectral changes in images to show its flexibility and superiority (Sun et al., 2022). In 2020, Ortis et al. (2019) sorted out the work of modern image sentiment analysis and analysed the development course of image sentiment analysis in the past decade (Khaki, 2024). They point out that image sentiment analysis has experienced three stages of development: low-level visual features, intermediate features and advanced learning features, and its method model has become mature at present. In this development process, researchers are no longer limited to extracting the feature representation of the image itself, but try to absorb information from other angles or patterns in an attempt to obtain sufficient basis for sentiment analysis. Therefore, the development of both text sentiment analysis and image sentiment analysis will eventually come to the same end.

Cross-modal fusion: in the natural social network environment, user data generally has the characteristics of randomness and abstractness, which causes great obstacles to mining the correlation between graphic patterns. The text of cross-media platforms aims to extract the original information of images and texts and build a unified representation of fusion features for subsequent processing and analysis (Hu et al., 2021). As the ultimate goal of intermodal fusion is to obtain the characteristic representation of the fusion signal, direct splicing of different modal features has become a standard fusion

method for early graphic sentiment analysis (Liu et al., 2021). Research has shown that the stitching operation fails to uncover deeper interactions between text and images. Therefore, they extract local and global features of images and calculate text and pictures based on attention mechanism matching to achieve cross-modal fusion (Liu et al., 2021).

The existing cross-modal fusion methods are no longer limited to simple splicing and information integration (Pathik and Shukla, 2022). According to the needs of practical problems and application environment, the influence of text and image on the whole system is analysed, the correlation between them is discussed, and the appropriate information is selected to construct the fusion feature representation. Graphic sentiment analysis is more in line with the needs of consumer sentiment analysis. Cross-modal fusion is an important core of multi-modal sentiment analysis. Contrastive language-image pretraining (CLIP) (Hafner et al., 2021) models embed images and text into the same space through contrast learning, but it mainly focuses on similarity matching between images and text, rather than emotion analysis tasks. In contrast, the model based on image translation is more focused on the extraction and analysis of emotional information, and can provide more specific consumer emotional information. Vision-and-language transformer (ViLT) (Kim et al., 2021) realises multi-modal tasks through the transformer architecture that directly processes images and texts. However, it does not eliminate redundant features in the processing of image features, which may lead to the problem of information redundancy. The model based on image translation can make more efficient use of image information through the previous feature selection step. Learning cross-modality encoder representations from transformers (LXMERT) (Tan and Bansal, 2019) adopts multi-level interaction mechanism when processing visual and linguistic tasks, but its complexity and computational cost are high. This model reduces the computational complexity by simplifying the feature input and enhancing the text data, while maintaining the accuracy of sentiment analysis.

There are fundamental differences in the structure of text and visual data, which pose challenges for feature alignment. Text is usually a linear sequence, while images are two-dimensional data. This structural difference complicates the direct comparison and fusion of the features of the two modes. Therefore, the algorithm proposed in this paper adopts a multi-level feature alignment method. By using a deep learning model for feature extraction of text and images, and using attention mechanisms to dynamically align the features of these two modes, the model is able to better understand and capture the correlations between different modes. Compared with the previous algorithms which only rely on the text sentiment analysis method, the proposed algorithm integrates the visual information in the image and converts the image into text description. Compared with other multi-modal sentiment analysis methods, the proposed translation method can better adapt to different types of images by using the emotion information in the images, accurately obtain the emotion information in the images, and form multi-modal input to enhance the comprehensiveness of sentiment analysis.

### 3 Methods

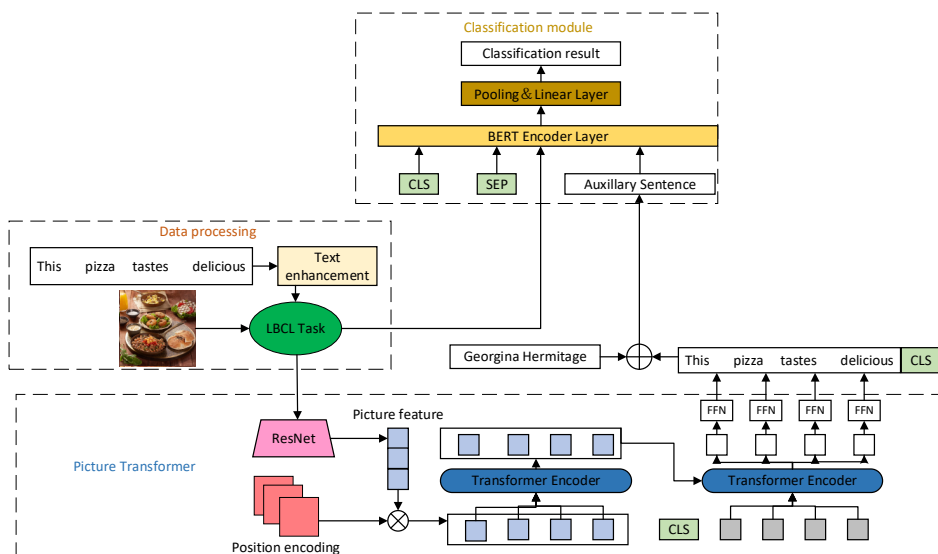
#### 3.1 Problem definition

Examples include images and text data. Each sample consists of text  $S_i = (w_1, w_2, \dots, w_n)$  with several words  $n$ . The output targets of text and images  $I_i$  are assigned a label – an

output target  $y_i \in \{\text{negative, neutral, positive}\}$ . The goal is to learn a function. In conclusion, imagine a tweet like “Crab is my favorite food, but I don’t like other seafood”, along with an image posted, the model to accurately predict the outcome or target. ‘Crab’ is cheerful, and target ‘seafood’ is negative.

The model proposed in this article consists of three parts, as shown in Figure 1. Given a multimodal input sample, it consists of an output target, an input sentence, and an image. In the data processing part, text data is first enhanced, and then contrast learning is carried out to help the model extract emotion-related features. Transformer converts images into vectors, transforming them from three-dimensional tensors of continuous data to one-dimensional integer vectors representing symbolic data. Finally, the classification part of the model, by sharing markers between the character converter and the large language model and mapping the symbol words to the vocabulary, effectively gives the image a natural language description for emotion classification.

**Figure 1** Overall structure of the model (see online version for colours)



### 3.2 Data processing

Label-based comparison learning method is employed to achieve emotional feature extraction from multimodal data for the model to achieve emotional feature extraction. As illustrated in Figure 1, the data contrastive learning task (ZCBZ). The algorithm significantly improves the effect of multi-modal emotion analysis by accurately processing emotion labels and generating unmasked labels, combined with the feature representation of MLF multi-layer fusion model. Method first divides each batch of data into positive and negative examples based on their emotional labels. A detailed algorithm can be found in Algorithm 1.

**Algorithm 1** ZCBZ

---

```

1  Start  $Z_c = [Z - 0, Z - 1, Z - 2]$  and  $Z_t = \text{list}()$ 
2  for  $i = 1; i < B; i++$  do
3      Start  $\bar{Z}_t = \text{list}()$ 
4      for  $j = 1; j \leq T; j++$  do
5          if  $Z_c[i][j]$  equals 0 then
6               $\bar{Z}_t \cdot \text{append}(j)$ 
7          end if
8      end for
9       $Z_t \cdot \text{append}(\bar{Z}_t)$ 
10 end for
11  $R = \text{MLF}(T, I)$ 
12  $\bar{Z}_{pn} = \text{einsum}(ns, ck \rightarrow nk, [R, R^T])$ 
13  $Z_{pn} = \text{LogSoftmax}(1_{pn}/\eta) \cdot \text{view}(-1)$ 
14  $Z_{cl} = Z_t[Z[1]]$ 
15 for  $q = 2; q \leq S; q++$  do
16      $Z_{cl} = \text{concat}(Z_{cl}, Z_t[Z[q]] + q \times T)$ 
17 end for
18  $Z_{zbcz} = \text{gather}(Z_{pn}, \text{index} = Z_{cl})/T$ 
19 return  $Z_{zbcz}$ 

```

---

Where the sentiment label is  $Z$ , batch a list of all the data. Let us say emotions are divided into three categories: positive (0), neutral (1), and negative (2). MLF’s multi-layer fusion model research text is  $T$  as follows: images are  $I$  the length  $Z_c B$  and the size of  $T$  the representation  $S$ . Einsum represents the Einstein summation convention, gather represents the aggregation of values with an indicator, and  $\eta$  represents the parameter of contrast learning. The algorithm mainly consists of two steps: first, the unmasking label is  $Z_t$  generated according to the data label in the batch processing, and then the loss matrix is  $Z_{pn}$  calculated to get the final loss.

We apply the method of data contrast learning to solve the problem of overfitting. By comparing textual and visual features of the same emotion category, the model is able to learn the internal associations between different modes, thereby enhancing its understanding of emotion. In data contrast learning, the introduction of unlabeled data is very important. Unlabelled data is generally richer and more pervasive than labelled data, providing more information and context to the model. The model is pre-trained on unlabeled data and potential features are extracted by contrast learning. These features can reflect the basic structure and distribution of data, and lay the foundation for subsequent supervised learning.

In order to improve the robustness of the model to the data and the learning ability of the model to recognise the invariant features in the data, we introduce a method based on

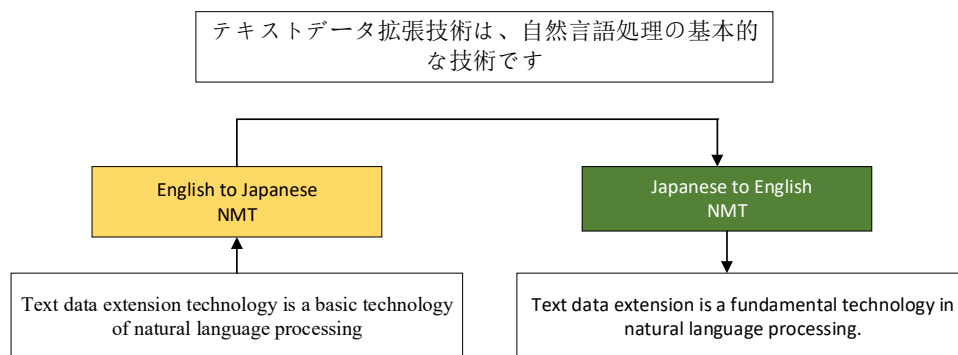


data enhancement. Due to the flexible expressiveness of text, traditional data enhancement can cause models to focus too much on the surface features of the data and ignore the effective features in the text. The emotional meaning expressed by the user should not change with the surface changes of the text. For example, “I ate bad pizza tonight.” And “I’m sorry” both convey negative meaning, “I’m sorry” is a valid feature, while other words change. Contrast learning based on data can help models better learn key features in the data and thus capture emotion-related features more effectively. This approach allows the model to make sense of the data and improve learning of emotion-related features.

In order to improve the performance and generalisation ability of the model, we can use various methods to deal with missing and noisy text data. This article uses translation techniques to enhance the text. Due to the significant progress in the field of text translation in recent years, text data enhancement based on inverse translation methods has become a high-quality text enhancement technology. Teams at CMU and Google Brain use back-translation as a specialised data augmentation technique to optimise the performance of question answering models (Yu et al., 2018). They trained two NMT models simultaneously, English to French and French to English, to achieve reverse translation, as shown in Figure 2.

The basic process of the translation is obvious. The translation model is to translate the original text of language A into language B, on the basis of the expression of language B into language C, and finally translate directly from the form of language C into language A, that is, the enhanced text of the original text. Let us take Google Translate for example. Raw text is a text data extension technology, which is an important technology in natural language processing. To be translated into Japanese: テキストデータ拡張技術は、自然言語処理の基本的な技術です; Japanese to English: natural language processing relies on the use of text data extension as a basic technique; As you can see, because the translation is good enough, the text before and after the enhancement remains semantically unchanged. Therefore, for the back translation enhancement technology, the quality of the translation model determines the final effect of data enhancement.

**Figure 2** Schematic diagram of back translation (see online version for colours)



Text enhancement plays an important role in multimodal emotion recognition, especially in solving the problem that some emotions are more common in the dataset, making it difficult for the model to handle less common emotions. Through text enhancement,

emotion expression in different contexts and scenes can be generated to help the model understand the changes of emotion in different contexts. This is especially important for uncommon emotions, as their expression may depend on the specific context. By enhancing uncommon emotions, the distribution of emotion categories in the dataset can be effectively balanced. This balance helps reduce bias in the model during training, making it more sensitive and accurate in the face of uncommon emotions.

### 3.3 Picture translation

The image translation task achieves efficient conversion from image to text by combining the structure of ResNet feature extraction, Transformer encoder and decoder. We translate the image  $I_i \in R^{3 \times H \times W}$  as an element  $\hat{I} \in N_0^l$  that, in the context of the large language model BERT (Yates et al., 2021), represents the symbolic natural language input. Considering the input image, a ResNet (Wu et al., 2019) is applied to generate an active map. The new feature map is flattened along the spatial dimension and obtained. After enhancement with fixed-position coding, the feature map is passed into the coding layer stack.

### 3.4 Transformer encoder

Considering the input image  $I \in R^{3 \times H_0 \times W_0}$ , ResNet is applied to show an active map; the new feature map is flattened along the spatial dimension and obtained. After enhancement with fixed-position coding, the feature map is passed into the coding layer stack.

The DETR encoder layer in DETR includes multiple attention heads that generate vital, query, and value embeddings.

$$[Q; K; V] = [T'_1(X_q + P_q); T'_2(X_{KV} + P_{KV}); T'_3 X_{KV}] \quad (1)$$

where  $P_q \in R^{d \times N_q}$  and  $P_{KV} \in R^{d \times N_{KV}}$  are the positions of the keys and queries embedded, respectively. In each transformer encoder layer, positional embeddings are added to help models understand the relationships between modes.  $T'_1$ ,  $T'_2$ ,  $T'_3$  constitute the weight  $T$ . The attention mechanism determines attention weights that reflect how important each element in the key-value sequence corresponds to the corresponding component of the query sequence. Set  $i$  as the query and the critical value  $j$ , then the calculation method of the weight of concern is  $\alpha_{i,j}$  as follows:

$$\alpha_{i,j} = \frac{e^{-\frac{1}{\sqrt{d'}} Q_i^T K_j}}{Z_i}, Z_i = \sum_{j=1}^{N_{KV}} e^{-\frac{1}{\sqrt{d'}} Q_i^T K_j} \quad (2)$$

The attention header first computes a weighted sum of values within a sequence, using attention as the weights. Subsequently, the attention heads from the encoder layer are combined into a multi-head attention by connecting the output of each attention head, as shown in equations (3) and (4):

$$(X_q, X_{KV}, T) = \sum_{j=1}^{N_{KV}} \alpha_{i,j} V_j. \quad (3)$$

$$X'_q = [\text{att}(X_q, X_{KV}, T_1); \dots; \text{att}(X_q, X_{KV}, T_M)] \quad (4)$$

The formula presented above is a representation of a key-query-value concern through the application of linear projection and residual connections.

### 3.5 Transformer decoder

The decoder transform size is  $D$ . The maximum sequence length that the transformer can handle. The system accepts object queries and utilises set matching losses to train the converter. The decoder is then employed in non-autoregressive text generation; this method generates descriptive text of the image according to the given image input by way of forward prediction. The generated description is:

$$v^* = \text{concat}(\text{TokenID}[\text{CLS}], \text{zeros}(l-1))$$

for the BERT token ID of the [CLS] token to join the dimensional zero vector. The resulting vector is the decoder's prompt, representing the beginning of the sentence. We encode to add a standard position, and the encoder receives a sequence of characters and returns an encoding of dimension  $D$ . In the decoder section, the RELU-activated fully connected layer is used to predict the most appropriate word for each position in the query from the vocabulary of the pre-trained BERT model. The tag is used to identify the end of the sentence and is usually inserted at the end of the sequence. For each position in the cue sequence (except for the first position that may contain a specific control code), the decoder generates a 30, 522-dimensional probability distribution representing the probability of each word appearing at that position, which reflects the model's prediction of the following word:

$$p(t[v^*, i]) = \text{SoftMax}(W_3 \times R(W_2 \times R(W_1 \times D_{ENC}(v^* i))) \quad (5)$$

where are the embeddings at the  $i^{\text{th}}$  position,  $W_3$ ,  $W_2$ ,  $W_1$  is the weight matrix learned, and is the ReLU activation function.

The training goal of the model is to optimise an objective function, which is the minimum of the sum of negative cross-entropy in each time step. To be specific:

$$L_1 = \min \sum_{i=1}^l -I(t_i) \log p(t_i = t_i^*) \quad (6)$$

where  $t_i^*$  is the probability distribution, which  $t$  is the correctly labelled single heat vector that appears in the coding title.

### 3.6 Classification module

Once we have trained the image translator, we can use it to translate the input image into a natural language description. With the auxiliary sentence mechanism employed, BERT

is used for the sentence pair classification pattern. In sentence pair classification mode, BERT’s input takes the form of sentence pairs, as follows:

$$[cls]t_1^A, t_2^A, \dots, t_{A.len}^A[SEP]t_1^B, t_2^B, \dots, t_{B.len}^B[PAD]\dots[PAD] \quad (7)$$

where  $t_1^A$  is the mark of sentence A, which typically embodies the text to be classified?  $t_i^B$  is the mark of sentence B, and it provides auxiliary information requisite for initiating BERT. Specifically, the tagging of the emotion target is interconnected to the tagging of the image description predicted by the title converter, thereby constructing a multimodal, comprehensive auxiliary sentence. The content text and auxiliary sentence are organised into a sentence-pair classification pattern, with sentence B serving as the auxiliary sentence. The resulting sequence is fed into the BERT, and the pooled output of the [CLS] token is utilised for further processing. Set  $H^{[CLS]} \in R^{768}$  as pool output, as follows:

$$p(y|H^{[CLS]}) = \text{softmax}(\theta_{Linear} \text{Dropout}(H^{[CLS]})) \quad (8)$$

fine-tune the formula (6) on the BERT encoder using the standard cross-entropy loss to get the formula (9):

$$L_2 = -\frac{1}{D} \sum_{j=1}^{|D|} \log p(y^j | H^{[CLS]j}) \quad (9)$$

Finally, the final classification results are output through the pooling layer and the linear layer.

## 4 Experimental setup

### 4.1 Dataset and parameter settings

The Twitter dataset consists of graphic messages posted on the social platform Twitter in 2014–2015 and 2016–2017. For both datasets, aspect terms now primarily refer to named entities (i.e., location, person, activity) in a social context. There are three types of emotional polarity for the Twitter dataset, namely positive, neutral, and hostile. The two datasets were combined into three datasets, including training, development, and test instances, at a ratio of 3:1:1. The dataset contains updated user comments and emotional expressions, reflecting changes and trends in emotional expression on social media in recent years. The dataset provides a strict emotional label to ensure the quality and reliability of the data, and provides a good basis for the training of the model. Twitter contains text and image data and supports multimodal sentiment analysis, which helps models better understand the relationship between text and images.

The Multi-ZOL dataset includes a randomly selected sample of mobile phone reviews, available at ZOL.com, and consists of a total of 5,288 multi-modal reviews, each consisting of a text, an image, and one to six emotions, each capable of being rated on a scale of 1-10. Each element is then divided into a triplet of emotions, corresponding text, and images. The final result is a total of 28,469 triples. The dataset is randomly divided into 80% training, 10% validation, and 10% testing. Datasets by combining data from different modes, models can provide a more comprehensive understanding of emotional expression. The dataset provides detailed affective labels, including multiple affective

categories such as positive, negative, and neutral, making the training and evaluation of the model more standardised and systematic. Including data from different domains (such as product reviews, social media, movie reviews, etc.) allows the model to learn the characteristics of emotional expression in different domains, thereby enhancing its ability to generalise.

It is reasonable to select Multi-ZOL, Twitter-2015 and Twitter-2017 datasets to study the multi-modal sentiment analysis model based on image translation. These three datasets provide a variety of emotional expressions and rich contextual information, which can help models learn more complex emotional patterns. These datasets are based on real user-generated content, making the model more relevant and valid in real-world applications.

Regarding privacy issues, do this by setting clear user agreements or privacy policies. When using comment data, try to remove or blur the user's personal information, including name, profile picture, etc. The accompanying pictures are blurred to protect the identity of the user. Collecting and analysing only the necessary information and avoiding the collection of personal data that is not relevant to the purpose of the analysis helps to reduce privacy risks.

During the training process, the Adam algorithm is used to optimise all model parameters to minimise cross-entropy loss. The implicit dimensionality of the model is set to 768 and the attention force is 8. In addition, in order to avoid overfitting, we use the dropout method to randomly discard some neural nodes with a dropout rate of 0.1. Finally, considering the differences between the Twitter dataset and Multi-ZOL, some hyperparameters take different values, including the number of training epochs, learning rate, and maximum word length. All models were implemented using PyTorch and trained on a GeForce RTX 3060 GPU.

## 4.2 Baseline model

CLIP (Zhou et al., 2022) uses learnable vectors to model the context of prompts and provides two model implementations for handling different image recognition tasks: unified context and class-specific context.

MAF (Xu et al., 2022) proposed approach introduces a general methodology for matching and calibrating, serving to bridge the technological gap between picture and text modes.

HVPNeT (Chen et al., 2022) proposed research construct an innovative layered visual prefix fusion network specifically designed to potentiate fusion representations.

ITA (Xu et al., 2022) aligns image features into the text space for better transformer embedding.

I2SRM (Wang et al., 2022) proposes a modelling approach for the in-sample and inter-sample relationships for this task.

TomBERT (Huang and Lin, 2023) is a target-oriented multi-model model, this model extracts the sensitive text and vision of the target based on the structure, and then integrates the multi-mode feature by stacking the self-attention layer stacking at the top.

### 4.3 Comparative experiments

To verify the effectiveness and advantage of our proposed MSST model, a comprehensive comparison experiment is carried out with many advanced algorithms in the current field. These algorithms include CLIP, MAF, HVPNeT, ITA, I2SRM, TomBERT, and so on. The experimental results are shown in Table 1. Through data analysis, it is seen that on three widely recognised databases, Multi-ZOL, Twitter-2015, and Twitter-2017, the MSST model shows excellent performance and the best effect.

Specifically, a detailed analysis of the experimental findings derived from the Multi-ZOL database is presented, and compared with the CLIP and TomBERT models, the accuracy rate of the MSST model is improved by 7.94% and 7.15%, respectively. The two models could be better at dealing with specific domains or complex emotional expressions because they lack targeted training and understanding of emotional contexts. The MSST model deals with emotion analysis through the perspective of picture translation, an innovation to the traditional multimodal fusion method, which helps to understand better and convey emotional information. On the Twitter-2015 database, the MSST model was compared with the ITA and MAF models. The results of the experiment showed an improvement in the MSST model's F1-score, which was 2.26% and 4.58% higher than the ITA and MAF models, respectively. This is because the ITA model's alignment of image features into the text space is insufficient to capture the deep interaction between images and text, especially for tasks that require deep semantic understanding, such as sentiment analysis. The MAF model's generic matching and calibration methods only partially adapt to some types of picture and text relationships when dealing with specific sentiment datasets. Analysis of experimental results fully proves the advantages of the introduction of contrast learning in the MSST model, which can better capture the fine-grained differences between pictures and texts by comparing the feature differences of different modes, thus improving the accuracy of sentiment analysis. The experimental data on the Twitter-2017 database show that in contrast to HVPNeT and 2SRM, the accuracy of the MSST model has been enhanced by 2.51% and 0.94%. These two models have limitations in fusion representation or relationship modelling and need to consider the particularity of sentiment analysis tasks fully. The study findings provide empirical evidence of the mighty power of the MSST model when dealing with large-scale, complex social media data.

In summary, through the comparative experiments on different databases, the MSST model not only achieves the best results in various performance indicators but also has a significant improvement effect compared with other advanced algorithms.

In this experiment, the accuracy of the model on different datasets was analysed in detail, including the calculation of confidence intervals and statistical significance tests. We calculate the following confidence intervals:

Multi-ZOL: [70.62, 73.26]; Twitter-2015: [78.13, 79.87]; Twitter-2017: [87.65, 89.05]. A confidence interval provides a range within which, at a certain confidence level (usually 95%), we can be 95% confident that the true accuracy falls. For the Multi-ZOL dataset, the confidence interval is [70.62, 73.26], indicating that we are confident that the true accuracy of the model on the Multi-ZOL dataset is likely to be between 70.62% and 73.26%. For the Twitter-2015 dataset, the confidence interval is [78.13, 79.87], indicating that the accuracy of the model on this dataset is high, possibly between 78.13% and 79.87%. For the Twitter-2017 dataset, the confidence interval is [87.65, 89.05], showing

that the accuracy of the model is the best on this dataset, and the true value is likely to be between 87.65% and 89.05%.

We conducted a statistical significance test, and the results showed a significant difference in accuracy between Multi-ZOL and Twitter-2015 ( $P < 0.01$ ). In statistical analysis, significance tests are used to determine whether the observed effect is likely to be caused by random error. We set the null hypothesis ( $H_0$ ) as “no significant difference in accuracy between Multi-ZOL and Twitter-2015”. By calculating the T-value and p-value, we find that the P-value is less than 0.01, which means that we have enough evidence to reject the null hypothesis for a significant difference in accuracy between the two. The accuracy of Multi-ZOL is significantly lower than that of Twitter-2015, which may indicate that the features of the Twitter-2015 dataset are better suited to the current model, or that Multi-ZOL has certain limitations in processing this dataset.

**Table 1** Comparative experiments of different models

	<i>Multi-ZOL</i>		<i>Twitter-2015</i>		<i>Twitter-2017</i>	
	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>
ITA	-	-	-	76.01	-	86.45
MAF	-	-	71.86	73.42	86.13	86.25
HVPNeT	-	-	73.87	75.32	85.84	86.87
CLIP	64.00	59.70	-	-	-	-
I 2SRM	-	-	76.22	77.12	87.41	88.40
TomBERT	64.79	64.27	77.15	71.75	70.34	68.03
MSST	71.94	70.45	79.00	78.27	88.35	89.86

#### 4.4 Ablation experiment

To gain a deeper understanding of the contribution and importance of each component in the MSST model, we designed and conducted three sets of ablation experiments. These experiments were divided into one set of single-mode experiments and one set of experiments that eliminated contrast learning, and the data were shown in Table 3.

In single-modal experiments, we focus on evaluating the performance of image emotion analysis to see how the model performs when relying only on a single mode. In the case of removing picture information and relying solely on text, we observed a significant decline in both the model’s recognition accuracy and F1-score.

Specifically, on the Multi-ZOL database, when only text data was used for emotion recognition, the model’s recognition accuracy decreased by 8.62%, and the F1-score decreased by 10.45%, recall fell 14.97. Similarly, when a similar experiment was conducted on the Twitter-2015 database, the accuracy dropped by 13.33%, and the F1-score declined by 9.38%. These data demonstrate the importance of the multimodal fusion approach to emotion recognition tasks, proving that the fusion of multimodal information such as text and pictures can significantly improve the performance of the model. To explore the effect of ZBCZ contrast learning on the model, we designed an experiment to delete the ZBCZ contrast learning task. The experimental results have demonstrated the following: in the Multi-ZOL database, once ZBCZ contrast learning is removed, the accuracy of the model is significantly reduced by 5.39%, and the F1-score is also reduced by 5.21%, recall fell 11.5%. This significant performance decline

highlights the core position of ZBCZ contrast learning in multimodal emotion recognition tasks, and evidence strongly corroborates its paramount contribution to the enhancement of the model’s performance.

**Table 2** Model ablation experiments

	<i>Multi-ZOL</i>			<i>Twitter-2015</i>		
	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>Recall (%)</i>	<i>ACC (%)</i>	<i>F1 (%)</i>	<i>Recall (%)</i>
No picture transformer	57.32	56.00	66.35	65.67	64.89	77.36
No ZBCZ	60.55	61.24	69.82	72.35	73.20	81.52
MSST	71.94	70.45	81.32	79.00	78.27	84.34

Through the comparative analysis of these ablation experiments, we can draw a clear conclusion: the importance of image data in consumer review datasets cannot be ignored. The efficacy of emotion recognition via multimodal fusion is markedly superior to that of single-modal approaches. To sum up, these experiments not only validate the necessity of multimodal fusion in emotion recognition tasks but also emphasise the importance of ZBCZ contrast learning in improving model performance. The MSST model achieves excellent performance in the field of multimodal sentiment analysis through effective integration of picture and text information, as well as well-designed contrastive learning mechanism.

#### 4.5 Visual analysis




Two models, MSST and BERT, were used to evaluate the accuracy of emotion recognition when conducting case studies. As shown in Table 3, three typical pictures and comments from Twitter in 2017 were selected for the study. The results reveal the limitations of relying on text analysis alone when it comes to understanding user emotions. In the second example, the content implied by the text needs to be understood in conjunction with the picture, which expresses humour and conveys a positive message. In the third example, the text appeared neutral on the surface, but in the context of the image, the actual emotion was positive. This suggests that relying solely on text analysis can lead to misunderstandings; while taking picture information into account, the utilisation of such data has the potential to enhance the accuracy of emotion prediction models significantly.

To verify the effectiveness of the contrastive learning task proposed in this study in learning emotion-related features, we conducted a visualisation experiment of multimodal data on a Multi-ZOL dataset. By dimensionally reducing the data feature vector output from the last layer of the model and applying the TSNE algorithm, we successfully mapped these features into a two-dimensional space for easy observation, as shown in Figure 3.

Figure 3(a) shows the feature distribution learned by the model when it is based on text data only and does not include contrast learning tasks. In contrast, Figure 3(b) shows how the model learns about emotional features when our proposed ZBCZ contrast learning task is not applied. Finally, Figure 3(c) shows the model’s comprehensive understanding and differentiation of emotional data after applying our proposed MSST model (incorporating the contrast learning task).



**Table 3** Specific case studies (see online version for colours)

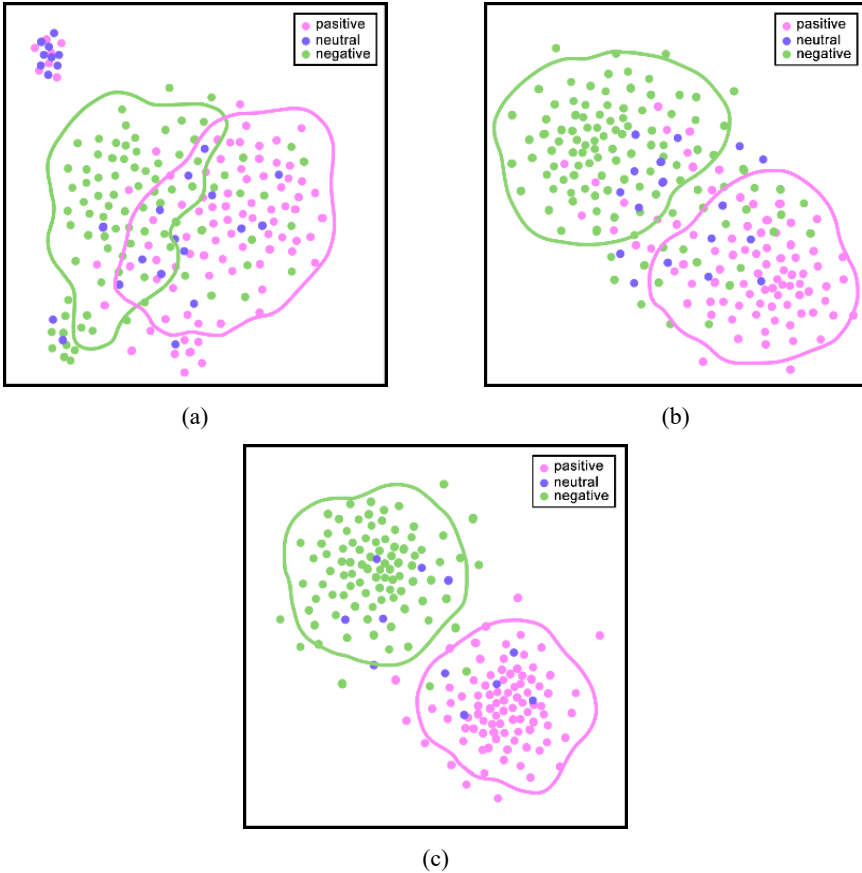
<i>Picture</i>	<i>Text</i>	<i>MSST</i>	<i>BERT</i>
	Mr. Krabs is under siege.	Negative	Negative
	The child brother's life is too bitter, suggested to hold the dog to my home, let me suffer such suffering!	Positive	Negative
	It looks so delicious. I want to eat it.	Positive	Neutral

From the visualised results, we can observe through Figures 3(a) and 3(c) that pink positive emotions are more clearly distinguished from green negative emotions in Figure 3 after the introduction of comparative learning tasks. This effect indicates that the model has learned how to extract and utilise key features in the emotional data and can effectively distinguish data with different emotional polarities. This distinction is based not only on simple statistical distributions, but also on a deep understanding of the nature of emotion. The effectiveness of the contrast learning task is that it forces the model to focus not only on the surface features of the data, but also on the standard features behind the data. These standard features are crucial to emotion analysis because they can capture the essence of emotion across different forms of expression.

As can be seen from the visualisation results of Figures 3(b) and 3(c), image translation plays a crucial role in emotion recognition. Compared with the multi-colour jumble of the original image [Figure 3(a)], the images processed by image translation [Figure 3(b) and 3(c)] are more clearly and accurately demarcated in neutral purple emotion, pink positive emotion, and green negative emotion. The core of image translation is to make it easier for the model to capture the emotional information contained in the image by converting the original image into a more precise and expressive form. Through image translation, we can remove noise and highlight the main expression of emotion, which is consistent with the simplified model. Classify the images into the correct emotional categories. With appropriate methods, the risk of misjudgement can be significantly reduced and the accuracy of the model in emotion recognition tasks can be improved.

In summary, the introduction of image translation and contrast learning tasks not only improves the model's performance on emotional analysis tasks, but also enhances the model's ability to understand emotional data. This improvement in understanding ability makes the model more comfortable in processing complex and changeable emotional data, providing strong technical support for the further development of the field of sentiment analysis.

**Figure 3** Visual analysis, (a) no ZBCZ (b) no picture transformer (c) MSST (see online version for colours)



## 5 Conclusions

With the boom in social networking and e-commerce, reviews have become an even more important part of websites. Instead of a single mode of text in the past, more and more consumers express their emotions through multi-mode data. It is an inevitable research trend to conduct sentiment analysis on multi-modal data. For image and text data, we proposed a multi-mode emotional analysis algorithm based on image translation. Use a transformer to convert the image into text in the input space and then use auxiliary sentence input encoders to integrate multimodal language models. Multiple

experiments show that image translation improves the fusion degree of picture and text in consumer reviews, retains the emotional information of pictures, and improves the accuracy of emotion recognition. Our proposed emotion recognition model may face the problem of classification error when dealing with emotion expression in different cultural backgrounds. To effectively address this problem, consider building future datasets that include emotional expressions from different cultural backgrounds. These datasets should cover multiple languages, customs, and cultural characteristics to ensure that models can learn a diverse range of emotional expressions. Cultural background annotation was introduced into the dataset to help the model understand the differences in emotional expression in different cultures. This can be done through expert tagging or crowdsourcing. In the feature selection process, cultural characteristics are considered to ensure that the model can capture the unique expression of emotions in different cultures.

## Acknowledgements

This research was supported by Humanities and Social Sciences Research Project of Chongqing Municipal Commission of Education (23SKGH465).

## References

- Bakhit, D.M.A., Nderu, L. and Ngunyi, A. (2024) 'A hybrid neural network model based on transfer learning for Arabic sentiment analysis of customer satisfaction', *Engineering Reports*, p.e12874, <https://doi.org/10.1002/eng2.12874>.
- Bening, S.A., Dachyar, M., Pratama, N.R., Park, J. and Chang, Y. (2023) 'E-commerce technologies adoption strategy selection in Indonesian SMEs using the decision-makers, technological, organizational and environmental (DTOE) framework', *Sustainability*, Vol. 15, No 12, p.9361.
- Bhatt, S., Bhatt, A. and Thanki, S. (2021) 'Analysing the key enablers of students' readiness for online learning: an interpretive structural modeling approach', *International Journal of Education and Development using Information and Communication Technology*, Vol. 17, No. 4, pp.105–130.
- Braghieri, L., Levy, R.E. and Makarin, A. (2022) 'Social media and mental health', *American Economic Review*, Vol. 112, No. 11, pp.3660–3693.
- Chauhan, P., Sharma, N. and Sikka, G. (2021) 'The emergence of social media data and sentiment analysis in election prediction', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 6, pp.2601–2627.
- Chen, X., Zhang, N.Y., Li, L., Yao, Y.Z., Deng, S.M., Tan, C.Q., Huang, F., Si, L. and Chen, H.J. (2022) *Good Visual Guidance Makes A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction*, CoRR abs/2205.03521.
- Dalmia, A., Gupta, M. and Varma, V. (2015) 'IIIT-H at SemEval 2015: Twitter sentiment analysis – the good, the bad and the neutral!', *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp.520–526.
- Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D. ... and Zavolan, M. (2021) 'CLIP and complementary methods', *Nature Reviews Methods Primers*, Vol. 1, No. 1, pp.1–23.
- Hou, M., Zhang, Z., Liu, C. and Lu, G. (2023) 'Semantic alignment network for multimodal emotion recognition', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 9, pp.5318–5329.

- Hu, J., Liu, Y., Zhao, J. and Jin, Q. (2021) *MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation*, arXiv preprint arXiv:2107.06779.
- Huang, Y. and Lin, Z. (2023) 'I2SRM: intra-and inter-sample relationship modeling for multimodal information extraction', *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, p.1.
- Jaiswal, A. and Arun, C.J. (2021) 'Potential of artificial intelligence for transformation of the education system in India', *International Journal of Education and Development using Information and Communication Technology*, Vol. 17, No. 1, pp.142–158.
- Khaki, A. (2024) 'Robust convolutional neural network based on UNet for iris segmentation', *International Journal of Image and Graphics*, Vol. 24, No. 4, p.2450042.
- Kim, W., Son, B. and Kim, I. (2021) 'Vilt: vision-and-language transformer without convolution or region supervision', *International Conference on Machine Learning*, PMLR, pp.5583–5594.
- Kode, H. and Barkana, B.D. (2023) 'Deep learning-and expert knowledge-based feature extraction and performance evaluation in breast histopathology images', *Cancers*, Vol. 15, No. 12, p.3075.
- Lei, Y. and Cao, H. (2023) 'Audio-visual emotion recognition with preference learning based on intended and multimodal perceived labels', *IEEE Transactions on Affective Computing*, Vol. 14, No. 4, pp.2954–2969.
- Liu, J., Chen, S., Wang, L., Liu, Z., Fu, Y., Guo, L. and Dang, J. (2021) 'Multimodal emotion recognition with capsule graph convolutional based representation fusion', *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.6339–6343.
- Liu, L.L., Yang, Y. and Wang, J. (2022) 'ABAFN: aspect-based sentiment analysis model for multimodal', *Journal of Computer Engineering & Applications*, Vol. 58, No. 10, pp.193–208.
- Mubeen, S., Kulkarni, N., Tanpoco, M.R., Kumar, R.D., Naidu, M.L. and Dhope, T. (2022) 'Linguistic based emotion detection from live social media data classification using metaheuristic deep learning techniques', *International Journal of Communication Networks and Information Security*, Vol. 14, No. 3, pp.176–186.
- Omuya, E.O., Okeyo, G. and Kimwele, M. (2023) 'Sentiment analysis on social media tweets using dimensionality reduction and natural language processing', *Engineering Reports*, Vol. 5, No. 3, p.e12579.
- Ortis, A., Farinella, G.M. and Battiato, S. (2019) 'Survey on visual sentiment analysis', *IET Image Processing*, Vol. 14, No. 8, pp.1440–1456.
- Pathik, N. and Shukla, P. (2022) 'Aspect based sentiment analysis of unlabeled reviews using linguistic rule based LDA', *Journal of Cases on Information Technology (JCIT)*, Vol. 24, No. 3, pp.1–19.
- Roohani, A. and Vinchek, M.H. (2023) 'Effect of game-based, social media, and classroom-based instruction on the learning of phrasal verbs', *Computer Assisted Language Learning*, Vol. 36, No. 3, pp.375–399.
- Sun, Y., Liu, B., Yu, X., Yu, A., Gao, K. and Ding, L. (2022) 'Perceiving spectral variation: unsupervised spectrum motion feature learning for hyperspectral image classification', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp.1–17, <https://doi.org/10.1109/TGRS.2022.3221534>.
- Tan, H. and Bansal, M. (2019) *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*, arXiv preprint arXiv:1908.07490.
- Tuncer, T., Dogan, S. and Subasi, A. (2022) 'LEDPatNet19: automated emotion recognition model based on nonlinear LED pattern feature extraction function using EEG signals', *Cognitive Neurodynamics*, August, Vol. 16, pp.1–12.
- Wang, X.Y., Gui, M., Jiang, Y., Jia, Z.X., Bach, N., Wang, T., Huang, Z.Q. and Tu, K.W. (2022) 'ITA: image-text alignments for multimodal named entity recognition', *NAACL-HLT. Association for Computational Linguistics*, pp.3176–3189.

- Wu, Z., Shen, C. and van den Hengel, A. (2019) 'Wider or deeper: revisiting the resnet model for visual recognition', *Pattern recognition*, Vol. 90, pp.119–133, <https://doi.org/10.1016/j.patcog.2019.01.006>.
- Xu, B., Huang, S.Z., Sha, C.F. and Wang, H.Y. (2022) 'MAF: a general matching and alignment framework for multimodal named entity recognition', *WSDM*, ACM, pp.1215–1223.
- Yang, H., Zhao, Y. and Qin, B. (2022) 'Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis', *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp.3324–3335.
- Yates, A., Nogueira, R. and Lin, J. (2021) 'Pretrained transformers for text ranking: BERT and beyond', *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp.1154–1156.
- Ying, Y., Yang, T. and Zhou, H. (2023) 'Multimodal fusion for Alzheimer's disease recognition', *Applied Intelligence*, Vol. 53, No. 12, pp.16029–16040.
- Yu, A.W., Dohan, D., Luong, M.T., Zhao, R., Chen, K., Norouzi, M. and Le, Q.V. (2018) *QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension*, arXiv preprint arXiv:1804.09541.
- Zhang, S., Ly, L., Mach, N. and Amaya, C. (2022) 'Topic modeling and sentiment analysis of yelp restaurant reviews', *International Journal of Information Systems in the Service Sector (IJISSS)*, Vol. 14, No. 1, pp.1–16.
- Zhang, X. (2022) 'Application of artificial intelligence in academic mental health and employment evaluation', *International Journal of Information Systems in the Service Sector (IJISSS)*, Vol. 14, No. 3, pp.1–15.
- Zhang, Y. and Zhang, L. (2022) 'Graphic and text emotional analysis based on deep fusion network', *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pp.955–958.
- Zhou, K., Yang, J., Loy, C.C. and Liu, Z. (2022) 'Learning to prompt for vision-language models', *International Journal of Computer Vision*, Vol. 130, No. 9, pp.2337–2348.