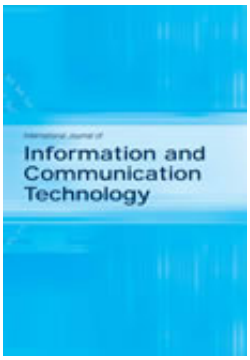
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

TSAIE-AppLing: multimodal sentiment analysis of image-enhanced text from a linguistic perspective

Huan Liu

DOI: 10.1504/IJICT.2025.10071316

Article History:

Received:	27 March 2025
Last revised:	09 April 2025
Accepted:	10 April 2025
Published online:	27 May 2025

TSAIE-AppLing: multimodal sentiment analysis of image-enhanced text from a linguistic perspective

Huan Liu

School of Literature and Journalism,
SiChuan University Jinjiang College,
MeiShan, 620000, China
Email: huanl1231@163.com

Abstract: To enhance the sentiment correlation between images and texts, this paper proposes a multimodal sentiment analysis approach for image-enhanced text from a linguistic perspective (TSAIE-AppLing). Firstly, bidirectional encoder representations from transformers (BERT) are introduced to encode textual features, and image features are extracted using visual transformer, which is combined with a multi-head self-attention mechanism to capture cross-modal global semantic features. Then we use null convolution to strengthen the feature association between image blocks and aggregate cross-block features, design a multi-head cross-attention mechanism to achieve inter-modal interaction alignment, use graph convolutional network (GCN) to enhance the textual semantic features related to the image, and carry out the final sentiment polarity determination through softmax function. Experimental results on the MVSA dataset show that the proposed method improves the classification accuracy by at least 2.75%, which can significantly improve the multimodal sentiment analysis.

Keywords: multimodal sentiment analysis; BERT model; multi-head cross attention; graph convolutional network.

Reference to this paper should be made as follows: Liu, H. (2025) 'TSAIE-AppLing: multimodal sentiment analysis of image-enhanced text from a linguistic perspective', *Int. J. Information and Communication Technology*, Vol. 26, No. 16, pp.69–84.

Biographical notes: Huan Liu received her Master's degree from SiChuan University in June 2015. She is currently working in the SiChuan University Jinjiang College. Her research interests include modern Chinese and natural language processing.

1 Introduction

The study of sentiment analysis can help us understand and explain human emotional expression, which has an important role to play in several areas of reality (Soleymani et al., 2017). With the booming of social media, people express their emotions by posting diverse posts on different topics on social media. Therefore, the use of social media data to analyse people's emotions has become a popular topic for researchers (Yue et al., 2019). In existing research on multimodal sentiment analysis, each pair of texts and images in a dataset is usually treated as a separate instance, without considering the

feature dependencies that may exist between different instances. Recently, researchers have found that social media posts have specific co-occurring features, such as co-occurring words, scenes, and objects, and these co-occurring features often have similar emotions (Chandrasekaran et al., 2021). By analysing texts from a linguistic perspective, we can explore lexical, syntactic, semantic, and other emotional features (Assem, 2022). Previous studies have looked for co-occurring features in picture modality, but we need to further explore the connection between picture and text modality through co-occurring features, so as to improve the accuracy and reliability of sentiment analysis.

Unimodal sentiment analysis is mainly focused on text and image domains. Constructing sentiment dictionaries was the dominant approach in early text sentiment analysis (Xu et al., 2019). Zucco et al. (2020) proposed the semantic orientation calculator (SO-CAL), which uses dictionaries of words annotated with semantic orientations, and is a dictionary-based approach to extract sentiment features from text. Zhang et al. (2018a) realised the sentiment analysis of microblogging netizen comment data by improving and expanding the sentiment dictionary, adding degree adverbs and negatives, etc., but it needs to consume more resources. With the growth of machine learning, text analysis methods have become more complex and accurate. Rhanoui et al. (2019) improved text classification accuracy by complementing CNN extraction of local features with bidirectional long-short-term memory (BiLSTM) extraction of global features. Li et al. (2022) combined BERT with BiLSTM to convert BERT hidden layer sequences into vectors to get semantic features input to BiLSTM, but the error of sentiment classification is large.

Earlier approaches to image sentiment analysis were mainly based on image sentiment classification methods based on low-end visual features. Khan et al. (2024) investigated the link between image features and sentiment by using visual features based on luminance, colour saturation, hue, and a combination of support vector regression and random forest. Yadav and Vishwakarma (2020) extracted image features specialised for the domain of artwork with emotional expressions and tested them on the International Affective Picture System (IAPS) to improve the classification results. Today's mainstream models for image sentiment classification are mainly based on deep learning models. Paolanti et al. (2019) improved the classification accuracy by changing the average pooling layer to double convolutional average pooling based on the ResNet network model. Cheng et al. (2024) performed image sentiment analysis based on the VGG19 migration learning method to achieve sentiment detection and classification of different emotions.

Due to the complexity and diversity of emotional expressions, it is often difficult to recognise emotions comprehensively and accurately using only a single modality. Therefore, combining information from multiple modalities for sentiment analysis improves the accuracy of sentiment recognition. Most approaches in multimodal sentiment analysis use both image and text modalities to obtain sentiment information. Gao et al. (2019) designed a goal-directed sentiment categorisation model based on the BERT architecture, which employs a goal-attention mechanism (AM) to achieve text and image alignment, further improving the accuracy of sentiment prediction. Aslam et al. (2023) extracted modal features through LSTM and multi-head attention to increase the weights of internally important features, and the evaluation metrics were all improved. Yadav and Vishwakarma (2023) used residual network to extract image features and used attention mechanism to obtain aspect-sensitive text and image features and validated the

effectiveness of the model on multiple datasets. Yang et al. (2020) proposed a multi-view attentional network sentiment analysis model for image and textual information, which improves the accuracy of sentiment recognition by continuously updating the memory network to obtain deep semantic features. Wang et al. (2024) proposed a fusion model of GCN and ResNet network, using GCN to effectively model the semantic relationship of text and ResNet network to enhance the image features, thus improving the accuracy of multimodal sentiment analysis.

From the above analysis of the current state of research, it can be seen that in the existing multimodal sentiment analysis tasks, the emotional correlation between images and text is neglected, resulting in a large amount of noise in the fused features, for this reason, this paper proposes the TSAIE-AppLing method. Firstly, BERT and BIGRU are utilised to encode the text features to enrich the contextual text feature representation, and visual transformer (ViT) is introduced to extract the image features that are similar to the text features. Subsequently, text and image descriptions are jointly encoded from a global perspective and combined with the multi-head self-attention mechanism (MSAM) to capture cross-modal global semantic features, from enabling the model to learn the interaction information between text and image modalities. Then, the graph structure is constructed from the local perspective to mine the fine-grained emotional information of text and images. The syntactic dependency graph is introduced into the text graph structure to enhance the text syntactic feature extraction. In the fusion map structure, hollow convolution is used to expand the sensory field to extract key information in the image blocks and strengthen the feature association across the blocks, and multi-head cross attention (MCCAM) is utilised to dynamically adjust the feature representations of the different modalities to achieve the inter-modal interaction alignment. Finally, the textual semantic related multimodal feature information is further enhanced by GCN which will be fused with image features and syntactic dependency matrix. Experimental results on MVSA-single and MVSA-multi datasets show that the classification accuracy of TSAIE-AppLing improves by 2.75%-13.25% compared to the benchmark model, with better sentiment analysis.

2 Relevant technologies

2.1 Graph convolutional network

GCN is a class of neural network models specialised for processing graph data, which extracts information about relationships and features between nodes through graph convolution operations (Levie et al., 2018). Compared to normal neural networks, GCN is able to operate directly on irregular graph structures without the need to convert them to regular lattice data. The core idea of GCN is neighbourhood aggregation, i.e., the representation of each node is not only determined by its own features, but is also updated through the interaction of information with its neighbouring nodes. The computation of the convolutional layer of GCN is shown below.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), H^{(0)} = X \tag{1}$$

where X is the constructed node feature and $\tilde{A} = A + I_N$ is the adjacency matrix, the self-cycling of the node feature is realised on the original graph by adding the unit matrix I_N to the original node similarity matrix A. $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the symmetric normalised Laplacian operator, which is used to normalise the adjacency matrix, $W^{(l)}$ is the learnable parameter matrix of the network, $\sigma(\cdot)$ is the activation function, and $H^{(l)}$ is the node characteristics of the l^{th} level.

2.2 Attention mechanism

AM is used for model learning and processing sequential data. It can be understood as people selectively focus their attention on the information they are interested in and ignore other irrelevant information (Soydaner, 2022). In the field of multimodal sentiment analysis, the core semantic information in the text and the regions in the image that express the sentiment become the information of interest to AM.AM can help the model to utilise the core information in the data effectively within the limited storage space and computational resources. For the obscured level sequence information $H = [h_1, h_2, \ldots, h_T]$, AM is feature extracted by weighted summation as follows.

$$y = \sum_{i=1}^{n} \alpha_i h_i \tag{2}$$

where h_i is the obscured information of moment i and α_i is the attention weight corresponding to h_i. The attention weight is transformed using softmax as follows.

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} e_j} \tag{3}$$

where e_i is computed by the matching function, the hidden layer represents the h_i output when the AM matching function is computed using a multilayer perceptron.

$$e_i = V_i \tanh(W_i h_i + b_i) \tag{4}$$

where W_i is the network parameters, tanh is the tangent activation function. h_i is the initial state of the obscured level, V_i, W_i are the weight coefficient matrices; b_i is the bias vector.

2.3 BERT model

BERT is a bi-directional encoder based on a transformer. The difference between BERT and GPT is that the transformer in the former utilises self-attention from both the left and right directions to learn the context in both directions, while the transformer in the latter is actually limited in self-attention, which results in the word only learning the context of the sequence from the left to the right. BERT obtains abstract bi-directional contextual features by masking the properties of the language model. In the pre-training phase, BERT is trained on a large-scale corpus to learn the generic features of the language (Min et al., 2023), mainly through masked language modelling (MLM) and next sentence prediction (NSP). The task of MLM is to randomly mask some words from the input sentence (i.e., replace them with a special [MASK] token), and then have the model

predict what these masked words are. NSP, on the other hand, predicts whether the given two sentences are consecutive texts. Both tasks can be automatically labelled from unsupervised data, and together they train BERT to capture rich linguistic features.

3 Text and image feature representation based on BERT and vision transformer

3.1 Text feature representation based on BERT and bidirectional GRU

Aiming at the heterogeneity issue of image and text data. In order to better fuse the features of text and image, the text features are first encoded using BERT and Bidirectional GRU (BiGRU) to enrich the contextual text feature representation. Then image features that are similar to text features are extracted using ViT, and finally text and image descriptions are jointly encoded from a global perspective and combined with the MSAM to capture cross-modal global semantic features, which emphasises the regions of common interest of text and image and enables the model to learn the interactive information of text modality and image modality. The text and image feature representation process is shown in Figure 1.

When the information in an image deviates from the content of the text, the textual description of the image enhances semantic clarity and thus reduces the linguistic gap between the two. This paper refers to the idea of CapTrBERT (Xiao et al., 2023), which utilises an image converter to convert the input image $I \in R^{3\times H\times W}$ into a textual description C of the image to provide textual level semantic information to help the model better understand the details and emotional expressions in the image.

First of all, this paper adopts the BERT model as a text feature extraction tool. Sentences and image descriptions are jointly encoded, and two special tagged nodes [*CLS*] and [*SEP*] are used to construct sentence pairs, and the input of the BERT coder is transformed into the form of '[*CLS*] + *sentence* + [*SEP*] + *image description*' as shown in equation (1).

$$X = \left\{[CLS], x_1^S, x_2^S, \ldots, x_{N_s}^S, [SEP], x_1^C, x_2^C, \ldots, x_{N_c}^C\right\} \tag{5}$$

where S is the input sentence and C is the image description, x_i^S and x_i^C are the i^{th} token of S and C, respectively, N_S and N_C are the lengths of the token sequences of S and C. Then X is fed into the BERT encoder to obtain the corresponding text-hidden representation $H_o = \{h_1, h_2, \ldots, h_{N_o}\}$, where H_o is the word vector of X, h_i is the BERT embedding of the i^{th} word, and N_o is the length of the input sequence.

To further capture the emotional information of the text, this paper adopts the BiGRU model to synthesise the pre- and post-textual information of the words so as to enhance the understanding of the overall meaning of the sentence.

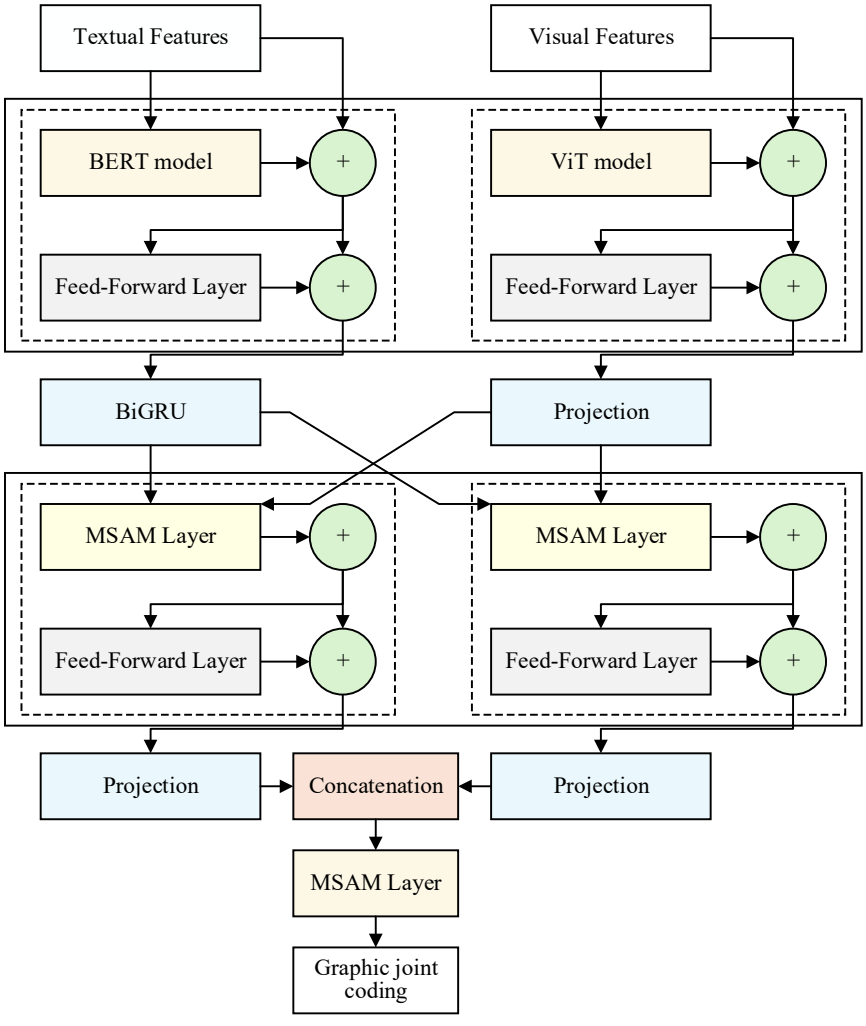
$$\vec{h}_i = GRU\left(x_i, \overrightarrow{h_{i-1}}\right) \tag{6}$$

$$\overleftarrow{h}_i = GRU\left(x_i, \overleftarrow{h_{i-1}}\right) \tag{7}$$

$$t_i = \frac{1}{2}\left(\overrightarrow{h_i} + \overleftarrow{h_i}\right) \tag{8}$$

where $\overrightarrow{h_i}$ is the forward state, $\overleftarrow{h_i}$ is the reverse state, and t_i is the i^{th} word feature. Finally, the word features are spliced together to obtain the word-level feature embedding encoding $T = [t_1, t_2, t_3, \ldots, t_n]$ of the text.

Figure 1 Multimodal feature representation (see online version for colours)

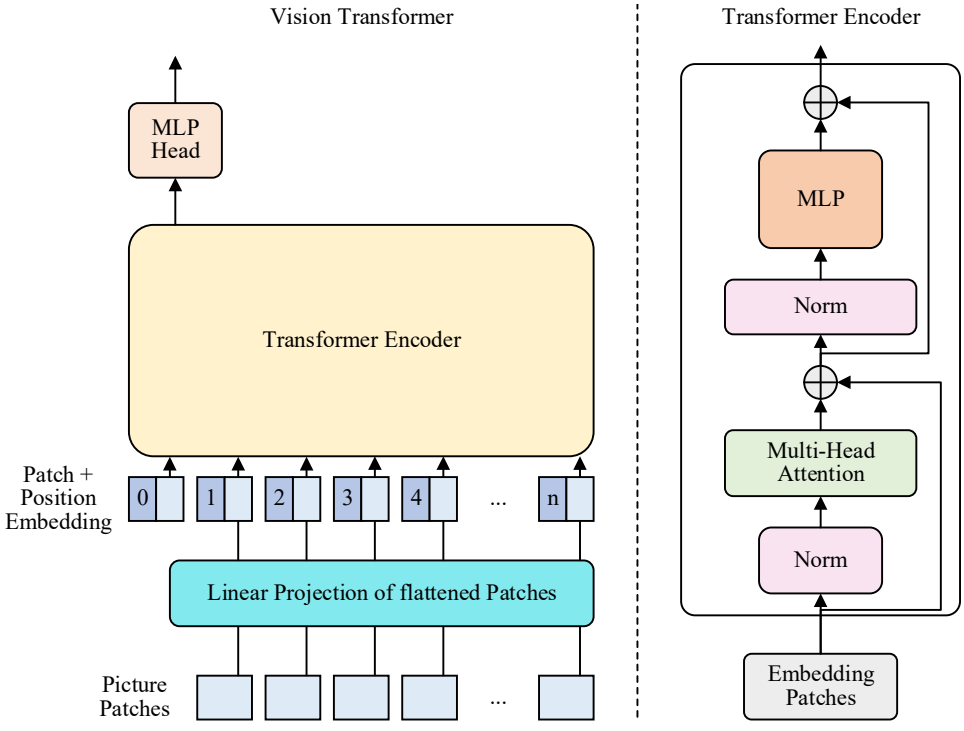


3.2 *Image feature coding based on ViT*

For image modal feature extraction, ViT (Naseer et al., 2021) is chosen as the core model for image feature extraction in this paper. Research scholars have experimentally demonstrated that ViT performs better compared to CNN and transformer. ViT achieves more powerful feature extraction capabilities than CNN and transformer with sufficient

data support through a global attention mechanism and convolutional-free design. It is particularly good at modelling long-range dependencies and complex feature interactions. However, its advantages are highly dependent on the scale of the data and computing resources, and in practical applications, it needs to be weighed and selected according to the requirements of the task. Using ViT it is possible to extract image features that are similar to text features, which facilitates subsequent fusion using cross-modal AM. In the process of processing, the image data is firstly processed in chunks, and then one-dimensional vector data is transformed by linear projection, and positional coding and flag bits are added to the sequence to better represent the information of whole data. The specific structure of ViT is shown in Figure 2.

Figure 2 ViT model (see online version for colours)



The input image is cropped, after this step the image is converted into n patch, the obtained patch is converted into a D-dimensional vector by linear transformation, the image features are extracted by using the ViT model, and the input image is divided into the image block $\{p_1, p_2, \ldots, p_{N_m}\}$ of $n \times m$, the sliced image block is flattened into a one-dimensional vector and goes through the linear projection level to obtain the final image representation H_m, as follows.

$$H_m = ViT\left(\{p_1, p_2, \ldots, p_{N_m}\}\right) \tag{9}$$

where N_m is the number of segmented image blocks.

3.3 Graphical joint coding based on multiple self-attention mechanism

To extract the global semantic information in text and image descriptions, this module calculates the global semantic correlation between text and image descriptions on the basis of joint representation combined with MSAM, so as to automatically adapt the semantic alignment between different modalities. MSAM utilises multiple attention heads to calculate the attention score of each word in context from different perspectives. The attention score of each word is obtained by calculating the weights of context words in the sentence, which ensures that the emotional features associated with the aspect words are effectively captured, enabling emotional features in the text and images to be characterised and analysed in a highly consistent manner, and the computational process is shown in equations (10)–(11).

$$W_i = \text{softmax}\left(\frac{[G_o W_q][G_o W_k]^T}{\sqrt{d_k}}\right) G_o W_v \tag{10}$$

$$G_{se} = concat\left(W_1, W_2, \ldots, W_j\right) W_c \tag{11}$$

where G_o is the joint feature representation of text and image, i.e., $T \oplus H_m$, W_q, W_k, W_v, W_q care the learnable weight matrices, W_i is the attention score of each head, d is the dimension of the key, $\sqrt{d_k}$ is the scaling factor, j is the number of attention heads, and G_{se} is the output after SAM.

4 Multimodal sentiment analysis of image-enhanced texts from the linguistic perspective

4.1 Syntactic dependency tree based representation of text grammars

Since the existing multimodal sentiment analysis methods ignore the semantic correlation between images and text, resulting in a large amount of noise in the fused features. For this reason, this paper proposes the TSAIE-AppLing method, as shown in Figure 3. Based on the joint feature coding of text and image, MSAM is combined to capture global semantic information and reduce the inter-modal semantic gap. Then, two graph structures are constructed to mine fine-grained sentiment information of text and images. Using null convolution to strengthen the feature association between image blocks and aggregate cross-block features, the MCCAM enhancement model is designed to focus on textual semantic features related to images, which solves the problem of inadequate extraction of text-related image features and improves the model's performance in sentiment classification.

Grammatical knowledge of sentences provides direct or indirect relationships between words, and by establishing links between words and aspectual words in a sentence through grammatical knowledge, aspectual words are better able to focus on the emotional information expressed by opinion words. Therefore, in this paper, text GCN is introduced to extract the deep feature information in the grammatical dependency graph to enhance text representation. First, the spaCy parser (Hu et al., 2022) is used to capture inter-word syntactic information to construct the syntactic dependency tree, and then, the

syntactic dependency tree is mapped to the adjacency matrix E, and the assignment process is shown as follows.

$$E_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \tag{12}$$

Second, GCN is utilised on the adjacency matrix to capture inter-word syntactic dependencies as follows.

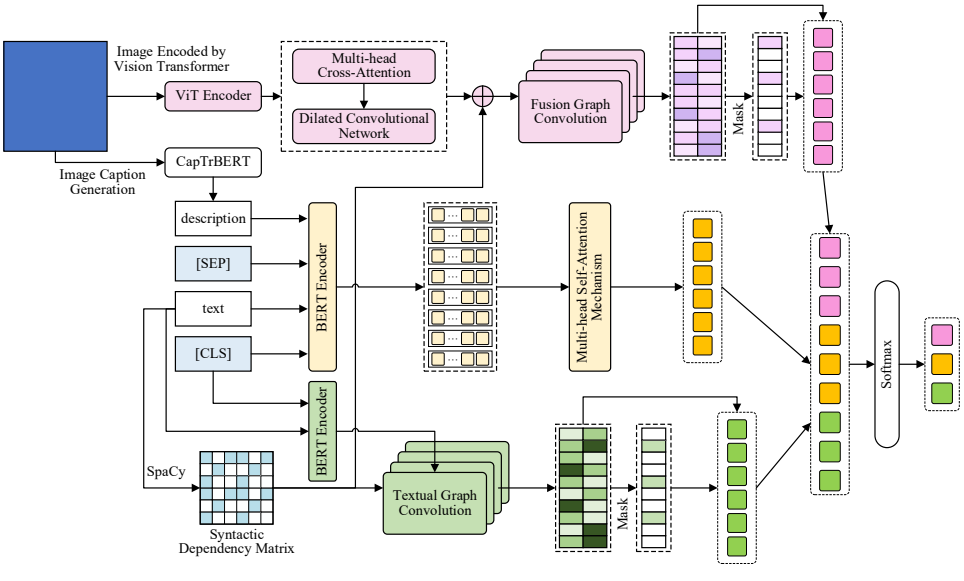
$$G_{sa,i}^{l} = ReLU\left(\frac{1}{d_i + 1}\sum_{j=1}^{n} E_{ij} W^{l} G_{sa,j}^{l-1} + b^{l}\right) \tag{13}$$

where E_{ij} is an element in E, $d_i = \sum_{j=1}^{n} E_{ij}$ is the degree of the i^{th} node in the syntactic dependency tree, W^l and b^l are the trainable weight matrix and bias term, respectively, $G_{sa,i}^{l}$ is the hidden representation of the i^{th} node in the l^{th} level, and $G_{sa,j}^{l-1}$ is the output of the previous network level. The output of TGCN after l level is G_{sh}.

$$M_i^{mask} = [0, \ldots, 1, \ldots, 0]^T \tag{14}$$

$$G_{sh} = \sum_{i=1}^{n} G_{sa,i}\left[M_i^{mask} G_{sa,i}\right] \tag{15}$$

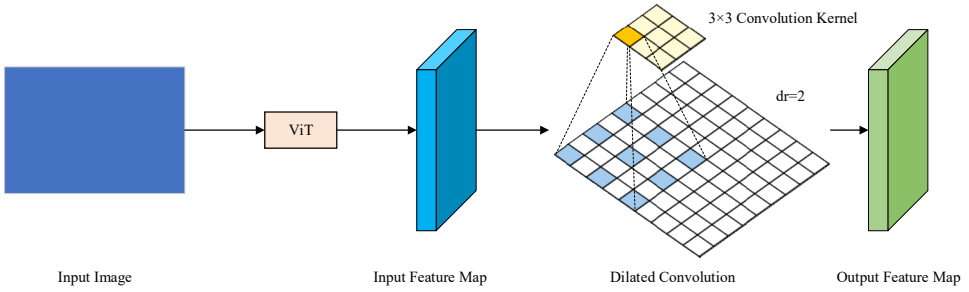
Figure 3 The structure of TSAIE-AppLing (see online version for colours)



4.2 Image semantic feature extraction based on null convolution

This paper introduces the cavity convolution (Zhang et al., 2018b) to further process the image. The use of cavity convolution expands the sensory field without changing the resolution or increasing the parameters, capturing a wider range of contextual information in the image, so that the model not only focuses on each individual image block, but also captures the correlation information across the blocks, which enhances the feature interaction of the chunks and effectively aggregates the information across the blocks.

Figure 4 Dilated convolution flowchart (see online version for colours)



In this paper, the convolutional kernel of 3 × 3 and the null convolutional network of $dr = 2$ are used to perform the convolutional operation on the ViT coded image features, as shown in Figure 4. In this paper, by inserting 1 hole between the neighboring elements of the convolutional kernel, the sampling distance of the convolutional kernel on the feature map is enlarged to 2, and thus the local features of the image are jump-sampled in order to capture a larger range of contextual information. The process of calculating the sensory field is as follows.

$$k' = k + (k-1)(dr-1) \tag{16}$$

where k' is the equivalent convolution kernel size mapped to the feature map, dr is the void ratio, and k is the actual convolution kernel size. The specific calculation of the void convolution is shown below.

$$G_{Conv,i} = \sum_{i=1}^{n} G_v[i + dr \times n] w[k'] \tag{17}$$

where G_{Conv} is the output eigenvalue after null convolution and G_v is the input.

4.3 Graphic modal alignment based on multiple heads of cross attention

To fully extract the image features related to aspectual words, this module designs MCCAM, which captures the multi-angle association between text and image through multiple attention heads, and uses the semantic information in the text to guide the model to focus on the image features related to aspectual words. First, the attention weights between text features and image features are calculated to measure the matching degree of the two modalities. Finally, the inter-modal interaction alignment is realised by

dynamically adjusting the feature representations of different modes. The computational procedure is shown as follows.

$$W_i^{'} = \text{softmax}\left(\frac{\left(G_{sh}\tilde{W}_q\right)^T\left(G_{Conv}\tilde{W}_k\right)}{\sqrt{d_k}}\right)\left[G_{Conv}\tilde{W}_v\right]^T \tag{18}$$

$$G_{lh} = concat\left(W_1^{'}, W_2^{'}, \ldots, W_j^{'}\right)W_c^{'} \tag{19}$$

where G_{sh} is the word vector of the sentence, G_{Conv} is the image features, $\tilde{W}_q, \tilde{W}_k, \tilde{W}_v$ and $W_c^{'}$ are the weight matrix and the linear transformation matrix of the output. $W_i^{'}$ is the attention score of each head, j' is the number of cross-attention heads, and G_{lh} is the final output of MCCAM.

After extracting the image features corresponding to the text, the text features are effectively fused with the image features as shown below.

$$G^f = LN\left(G_{se}, G_{lh}\right) \tag{20}$$

where G^f is the fusion feature, n is the input sequence length, d_f is the feature hidden layer dimension, and LN is the layer normalisation operation.

4.4 Image enhanced text and sentiment classification based on graph convolutional network

The semantically relevant multimodal feature information is further enhanced by fusing with image features and syntactic dependency matrix E. Specifically, the cosine similarity function is used to calculate the similarity between each node in the fused feature G^f, and the similarity matrix U is formed by combining the similarities of each two nodes. The Hadamard product operation is done on E and U and the adjacency matrix H is obtained by L2 normalisation as follows.

$$H = \|E \circ U\|_2 \tag{21}$$

The interaction between nodes is realised by passing the adjacency matrix H into the fusion GCN as follows.

$$G_{sy,i}^l = ReLU\left(\frac{1}{d_i + 1}\sum_{j=1}^{n} H_{ij}\hat{W}^l G_{sy,j}^{l-1} + \hat{b}^l\right) \tag{22}$$

where H_{ij} is the element in H, $d_i = \sum_{j=1}^{n} H_{ij}$ is the degree of the i^{th} node, $\hat{W}^l$ is the weight matrix of the l^{th} level, $\hat{b}^l$ is the bias term, $G_{sy,i}^l$ is the hidden representation of the i^{th} node in the l^{th} level, and $G_{sy,j}^{l-1}$ is the output of the previous GCN level. The richer image-enhanced text fusion feature G_{sy} is obtained after l-level graph convolution operation.

The semantic feature G_{se}, syntactic feature G_{sh} and graphic fusion feature G_{sy} are spliced to obtain the final representation of multimodal sentiment feature y, and the final

sentiment polarity determination is performed by softmax function, and the computation process is shown as follows.

$$\hat{y} = \text{softmax}\left(W_o\left[G_{se}, G_{sg}, G_{sy}\right] + b_o\right) \tag{23}$$

where W_o and b_o are the weight term and bias term of the fully connected layer, respectively. This paper uses the minimised cross-entropy loss function to optimise the model parameters, as shown in equation (24).

$$Loss = \sum_{i}^{\bar{S}}\sum_{j}^{\bar{C}} y_i^j \ln \hat{y}_i^j + \lambda \|\theta\|_2 \tag{24}$$

where S is the number of training samples, C is the number of categories, y_i^j and $\hat{y}_i^j$ represent the true and predicted labels of the training set, respectively, λ is the L2 regularisation term coefficient, and θ is all trainable parameters.

5 Experimental results and analyses

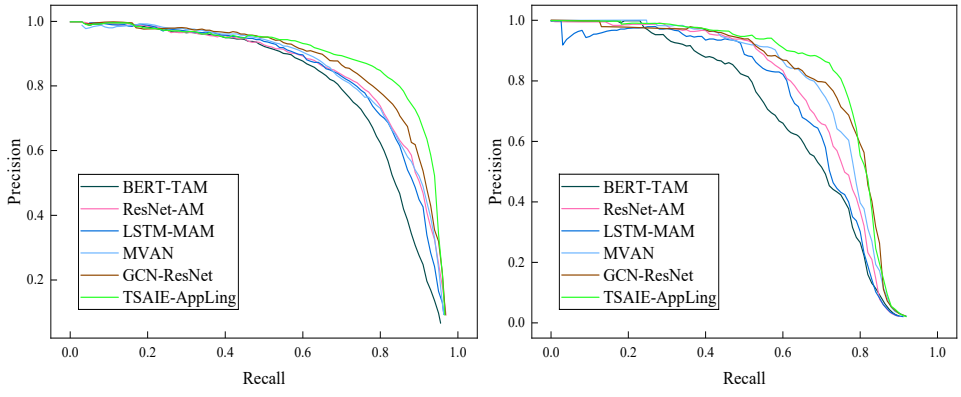
To validate the effectiveness of TSAIE-AppLing, this paper uses the MVSA dataset, which contains two independent datasets MVSA-single and MVSA-multi. This dataset collects image and text comment data from Twitter. The MVSA-single dataset contains 5,129 image-texts and corresponding sentiment labels, while the MVSA-multi dataset contains 17,024 image-texts and each image-text corresponds to three sentiment labels. The experiments validate the baseline model and the proposed method using Pytorch deep learning framework based on Python 3.8, CUDA version 11.1, Pytorch version 1.9.1, and RTX3090 graphics card with 24 GB of GPU memory, and train the model on a Linux server. In the experiments, the MVSA-single and MVSA-single datasets are divided into training set, validation set and test set according to the ratio of 8:1:1. The detailed information of datasets is shown in Table 1.

Table 1 MVSA dataset segmentation

Dataset	*Training set*	*Validation set*	*Test set*
MVSA-single	3,609	451	451
MVSA-multi	13,620	1,702	1,702

This paper performs comparative experiments on the baseline models BERT-TAM, LSTM-MAM, ResNet-AM, MVAN, GCN-ResNet and TSAIE-AppLing using PR curves, accuracy (A), F1 values, and mean absolute error (MAE). Comparison of PR curves of different sentiment classification methods on the two datasets is shown in Figure 5. By focusing on precision and recall, the PR curve can more objectively assess the model's ability to recognise a few categories (e.g., negative emotions). As can be seen from Figure 6, on the MVSA-single and MVSA-multi datasets, the PR curve areas of TSAIE-AppLing are 0.9765 and 0.9917, respectively, which are significantly higher than the comparison models, indicating that the classification accuracy of TSAIE-AppLing is better than the other five methods on each sentiment category.

Figure 5 PR curves of different sentiment classification methods on the two datasets, (a) PR curves on the MVSA-single dataset (b) PR curves on the MVSA-multi dataset (see online version for colours)



Comparison of A, F1, and MAE metrics of different methods on MVSA-single and MVSA-multi datasets are shown in Table 2. The A of TSAIE-AppLing on MVSA-single and MVSA-multi datasets is 92.63% and 95.33%, respectively, which is an improvement of 2.75%-13.25% compared to the benchmark model. The F1 of TSAIE-AppLing is 93.64% and 94.16%, respectively, which improves at least 2.61% compared to the benchmark model. The MAE of TSAIE-AppLing is 0.1318 and 0.0639, respectively, which reduces at least 12.66% compared to the benchmark model.

BERT-TAM requires pre-training on a large-scale corpus to learn an effective language representation. For low-resource languages or domain-specific tasks, the method may not perform as well as expected. LSTM-MAM is a sequential model where the computation of each step depends on the output of the previous step, making it difficult to fully utilise the parallel computing power of GPUs. Although ResNet-AM deeply mines graphical features, it does not consider inter-modal alignment and interaction, so the graphical feature fusion is not effective. MVAN mines semantic features of graphs through GAT, but does not address the cross-modal semantic divide. GCN-ResNet implements multimodal sentiment analysis of graphs through GCN and ResNet, but does not augment the semantics of graphs, and thus the classification accuracy is not as good as TSAIE-AppLing.

Table 2 Comparison of accuracy, F1, and MAE metrics

Method	*MVSA-single*			*MVSA-multi*		
	A (%)	*F1 (%)*	*MAE*	*A (%)*	*F1 (%)*	*MAE*
BERT-TAM	80.39	81.78	0.3166	82.08	82.72	0.2142
LSTM-MAM	82.02	82.59	0.2594	84.39	83.51	0.1728
ResNet-AM	83.51	84.46	0.2103	86.12	87.63	0.1589
MVAN	87.05	86.29	0.1927	89.84	89.01	0.1138
GCN-ResNet	89.48	91.03	0.1509	92.58	91.57	0.0926
TSAIE-AppLing	92.63	93.64	0.1318	95.33	94.16	0.0639

To visualise the advantages of TSAIE-AppLing, this paper uses the t-SNE method to visualise the MCCAM-based graphical modal alignment module (GMA-MCCAM), GCN-based image enhanced text fusion module (GCN-TF), and the learned sentiment features of TSAIE-AppLing.

Figure 6 Results of ablation experiments with different modules, (a) GMA-MCCAM (b) GCN-TF (c) TSAIE-AppLing (see online version for colours)

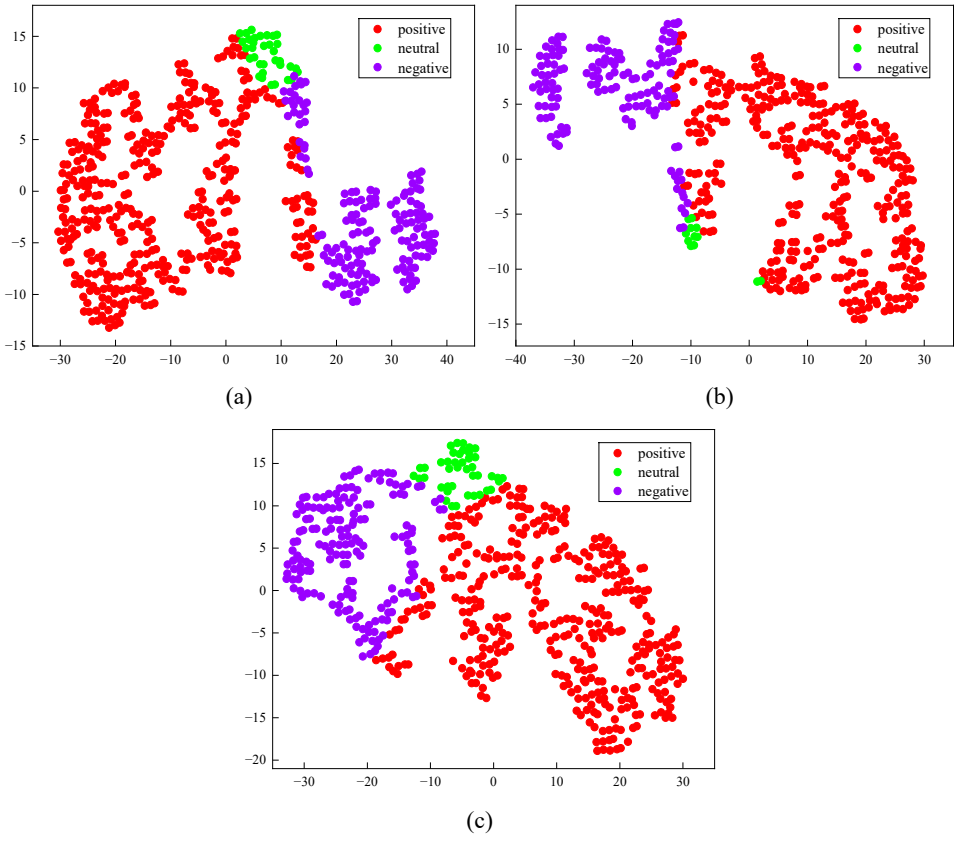


Figure 6 shows the visualisation results of the three modules on the MVSA-single dataset. Compared with GMA-MCCAM module and GCN-TF module, TSAIE-AppLing has more discriminative labels of emotional features on the dataset. Specifically, the distribution of the three sentiment labels in TSAIE-AppLing is more discriminative, and the same label points are still relatively concentrated after dimensionality reduction. This is because in TSAIE-AppLing, the interactions and influences of the emotional features of the modules help to define clearer data structures and boundaries, which may not be captured by a single module. This finding suggests that TSAIE-AppLing, which combines the above two modules, produces better classification results.

6 Conclusions

It is of great practical significance to study the sentiment tendency of graphic data, but the inter-modal differences and various problems in real data increase the difficulty for sentiment analysis research. In this regard, this paper proposes the TSAIE-AppLing method. Firstly, BERT and BIGRU are used to encode the text features, ViT is introduced to extract the image features that are similar to the text features, and then the text and image descriptions are jointly encoded from the global perspective, and MSAM is combined to capture the cross-modal global semantic features, from which the model learns the multimodal interaction information. Next, graph structures are constructed to mine fine-grained sentiment information of text and images. Null convolution is utilised to strengthen feature associations between image blocks and aggregate cross-block features, followed by the design of MCCAM to dynamically adjust the feature representations of different modalities to achieve inter-modal interactive alignment. The textual semantic related multimodal feature information is further enhanced by GCN which will be fused with image features and syntactic dependency matrix. Finally, multimodal sentiment polarity determination is performed by softmax function. Simulation experiments were conducted on MVSA-single and MVSA-multi datasets, and the results show that the classification accuracy of TSAIE-AppLing is improved by 2.75%–13.25% compared with the benchmark model, which has significant sentiment classification advantages. In the future research, TSAIE-AppLing will be improved, and it is planned to enhance the feature representation by introducing corresponding domain knowledge, fully mining the information closely related to the text from each modality, and further enhancing the multimodal feature interaction.

Declarations

All authors declare that they have no conflicts of interest.

References

Aslam, A., Sargano, A.B. and Habib, Z. (2023) 'Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks', *Applied Soft Computing*, Vol. 144, p.110494.

Assem, S. (2022) 'The development of sentiment analysis from a linguistic perspective', *The Egyptian Journal of Language Engineering*, Vol. 9, No. 2, pp.40–52.

Chandrasekaran, G., Nguyen, T.N. and Hemanth D, J. (2021) 'Multimodal sentimental analysis for social media applications: a comprehensive review', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 11, No. 5, p.e1415.

Cheng, J., Yang, L. and Tong, S. (2024) 'Painting style and sentiment recognition using multi-feature fusion and style migration techniques', *Informatica*, Vol. 48, No. 21, pp.127–138.

Gao, Z., Feng, A., Song, X. and Wu, X. (2019) 'Target-dependent sentiment classification with BERT', *IEEE Access*, Vol. 7, pp.154290–154299.

Hu, C., Gong, H. and He, Y. (2022) 'Data driven identification of international cutting edge science and technologies using SpaCy', *PloS One*, Vol. 17, No. 10, p.e0275872.

Khan, T.A., Sadiq, R., Shahid, Z., Alam, M.M. and Su'ud, M.B.M. (2024) 'Sentiment analysis using support vector machine and random forest', *Journal of Informatics and Web Engineering*, Vol. 3, No. 1, pp.67–75.

Levie, R., Monti, F., Bresson, X. and Bronstein, M.M. (2018) 'Cayleynets: graph convolutional neural networks with complex rational spectral filters', *IEEE Transactions on Signal Processing*, Vol. 67, No. 1, pp.97–109.

Li, X., Lei, Y. and Ji, S. (2022) 'BERT-and BiLSTM-based sentiment analysis of online Chinese buzzwords', *Future Internet*, Vol. 14, No. 11, p.332.

Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I. and Roth, D. (2023) 'Recent advances in natural language processing via large pre-trained language models: a survey', *ACM Computing Surveys*, Vol. 56, No. 2, pp.1–40.

Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F. and Yang, M-H. (2021) 'Intriguing properties of vision transformers', *Advances in Neural Information Processing Systems*, Vol. 34, pp.23296–23308.

Paolanti, M., Pierdicca, R., Martini, M., Felicetti, A., Malinverni, E., Frontoni, E. and Zingaretti, P. (2019) 'Deep convolutional neural networks for sentiment analysis of cultural heritage', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 42, pp.871–878.

Rhanoui, M., Mikram, M., Yousfi, S. and Barzali, S. (2019) 'A CNN-BiLSTM model for document-level sentiment analysis', *Machine Learning and Knowledge Extraction*, Vol. 1, No. 3, pp.832–847.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S-F. and Pantic, M. (2017) 'A survey of multimodal sentiment analysis', *Image and Vision Computing*, Vol. 65, pp.3–14.

Soydaner, D. (2022) 'Attention mechanism in neural networks: where it comes and where it goes', *Neural Computing and Applications*, Vol. 34, No. 16, pp.13371–13385.

Wang, H., Ren, C. and Yu, Z. (2024) 'Multimodal sentiment analysis based on cross-instance graph neural networks', *Applied Intelligence*, Vol. 54, No. 4, pp.3403–3416.

Xiao, L., Wu, X., Yang, S., Xu, J., Zhou, J. and He, L. (2023) 'Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis', *Information Processing & Management*, Vol. 60, No. 6, p.103508.

Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y. and Wu, X. (2019) 'Chinese text sentiment analysis based on extended sentiment dictionary', *IEEE Access*, Vol. 7, pp.43749–43762.

Yadav, A. and Vishwakarma, D.K. (2020) 'A deep learning architecture of RA-DLNet for visual sentiment analysis', *Multimedia Systems*, Vol. 26, No. 4, pp.431–451.

Yadav, A. and Vishwakarma, D.K. (2023) 'A deep multi-level attentive network for multimodal sentiment analysis', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 19, No. 1, pp.1–19.

Yang, X., Feng, S., Wang, D. and Zhang, Y. (2020) 'Image-text multimodal emotion classification via multi-view attentional network', *IEEE Transactions on Multimedia*, Vol. 23, pp.4014–4026.

Yue, L., Chen, W., Li, X., Zuo, W. and Yin, M. (2019) 'A survey of sentiment analysis in social media', *Knowledge and Information Systems*, Vol. 60, pp.617–663.

Zhang, S., Wei, Z., Wang, Y. and Liao, T. (2018a) 'Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary', *Future Generation Computer Systems*, Vol. 81, pp.395–403.

Zhang, Z., Wang, X. and Jung, C. (2018b) 'DCSR: dilated convolutions for single image super-resolution', *IEEE Transactions on Image Processing*, Vol. 28, No. 4, pp.1625–1635.

Zucco, C., Calabrese, B., Agapito, G., Guzzi, P.H. and Cannataro, M. (2020) 'Sentiment analysis for mining texts and social networks data: methods and tools', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 10, No. 1, p.e1333.