Advancing Depression Detection in Social Media: A Multimodal Aspect-Level Sentiment Analysis Approach

Yuxin Liu

School of Computer Science and Technology, Guizhou University, Guiyang, China

sdc.yxliu21@gzu.edu.cn

Abstract. Depression, a prevalent mental illness, increasingly manifests in the digital expressions of individuals on social media platforms. This paper explores the use of multimodal aspect-level sentiment analysis for detecting depressive tendencies from social media data, a method that surpasses traditional unimodal approaches in granularity and adaptability. By integrating textual and visual cues from users' posts, this paper's approach employs the Target-Oriented Multimodal Bidirectional Encoder Representations from Transformers (TOM-BERT) framework. This deep learning model is fine-tuned to discern subtle indicators of depression by analyzing the interplay between different types of data inputs. This paper's experimental setup compares this method against conventional models primarily focused on single-mode data analysis, demonstrating its superior capability in identifying depressive signals. Results reveal that this paper's multimodal approach not only captures a richer spectrum of emotional expressions but also enhances the accuracy of depression detection. This research underscores the potential of advanced sentiment analysis techniques in mental health monitoring, particularly in leveraging the nuanced data available through social networks.

Keywords: Depression Detection, Multimodal Aspect-Based Sentiment Analysis, Deep Learning.

1 Introduction

Depression is a significant public health concern in modern society, affecting millions of individuals and posing serious challenges to mental health care systems globally. It can result in long-term emotional distress, cognitive impairment, and slow thinking, among other issues, making it one of the most common and harmful psychological conditions. Early detection and treatment of depression are crucial. However, current diagnostic methods primarily rely on patients' self-assessment and the manual completion of questionnaires. This approach is not only time-consuming but also limited by the accuracy of patients' recollection of mood changes. Additionally, some patients may be reluctant to share their emotional state due to concerns about privacy or psychological resistance. Therefore, it is essential to explore alternative channels for effectively detecting depression automatically.

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_44

With the continuous development of social media platforms, their use has become integrated into every aspect of people's lives. An increasing number of individuals use social media to express themselves, and the data generated on these platforms often reflects their genuine and diverse emotions, offering new avenues for detecting depressive tendencies. Sentiment analysis is a critical task within the field of natural language processing, aimed at identifying and extracting emotional tendencies and attitudes from text. In recent years, using sentiment analysis technology for depression detection based on social media data has become an important research direction. Researchers have extracted specific features through machine learning and deep learning, which can help social media users identify their own mental health status [1]. Depression tendency detection based on social media platform data provides a convenient and rapid method for identifying depressive tendencies, holding strong practical value.

Choudhury et al. pioneered the use of social platform data for detecting depression tendencies, investigating the feasibility of this approach. Following Choudhury's research, some scholars began annotating social media datasets, making automatic depression tendency detection possible [2]. Various machine learning techniques have been applied to this task, including Support Vector Machines (SVM), decision trees, and random forest algorithms. Neural networks have also been utilized to detect users' psychological stress [3]. As deep learning has advanced, models like Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) have increasingly been utilized in sentiment analysis research, introducing new avenues for improving depression detection methods. Recently, pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT), XLNET, and A Robustly Optimized BERT Pretraining Approach (RoBERTa) have demonstrated exceptional performance in detecting depression tendencies.

Initially, researchers used text alone as input to detect depressive tendencies. However, users often express emotions through a combination of images and text rather than text alone. Consequently, depression detection based on multimodal sentiment analysis has emerged, aiming to fuse emotional information from both text and image modalities. Multimodal sentiment analysis can obtain more comprehensive emotional information from multiple perceptual channels compared to unimodal sentiment analysis. For instance, in a data sample combining text and image, the text might state, "The pizza at this restaurant is quite good," conveying a positive attitude, whereas if the accompanying image is negative, the overall sentiment of the data is likely to be negative. Considering only the text may yield a positive sentiment, but when combined with the image, it reveals a sarcastic negative sentiment [4]. In 2019, Rodrigues Makiuchi et al. proposed a CNN-BERT fusion multimodal depression detection model; in 2020, Chun et al. used Instagram datasets for multimodal depression detection [5, 6]. The fusion of multimodal data for depression detection is currently a research trend. While multimodal data offers richer information, effectively processing this information remains a critical challenge. Sentiment analysis can be categorized based on granularity into three levels: document-level, sentence-level, and aspect-level. Among these, aspect-level sentiment analysis stands out as a detailed task that evaluates emotional tendencies from different aspects

within a piece of text. Multimodal aspect-based sentiment analysis improves the accuracy of predicting sentiment tendencies by incorporating various modalities and aspects. This approach enables fine-grained processing and better integration of different types of data, leading to more comprehensive and precise sentiment analysis [7]. Fig. 1 illustrates an instance of multimodal aspect-based sentiment analysis, demonstrating how various modalities and aspects are integrated to analyze sentiment tendencies.

RT @ nbcsandiego : 13 - year - old boy missing in **Chula Vista**[neutral] . RT to spread the word .

RT @ BBCOne : Dear **Madonna**[positive] , THIS is how you wear a cape . # **Poldark**[neutral] # **Demelza**[neutral]

Fig. 1. An example of multimodal aspect-based sentiment analysis [7].

As a fine-grained task of sentiment analysis, the goal of multimodal aspect-level sentiment analysis is to identify and analyze the emotional tendencies associated with specific aspects within multimodal reviews, integrating data from various sources. This encompasses not only textual content but also other modalities such as images and audio. By integrating data from these various modalities, researchers can accurately capture and understand users' emotional attitudes from a particular aspect. This approach provides a more comprehensive reflection of the reviewers' emotional states, thereby enhancing the accuracy and practicality of sentiment analysis. It mainly revolves around two subtasks: multimodal sentiment classification based on aspect words and multimodal sentiment recognition based on aspect categorization [8]. Xu et al. proposed the MIMN model to coordinate textual input with visual input labeled with aspect words; Zhou et al. presented an algorithmic framework based on a unified framework for multimodal aspect word extraction and aspect-level sentiment classification; Ling et al. introduced a pre-training framework for multimodal aspect-level sentiment analysis, utilizing a BART encoder-decoder model specifically designed for task-oriented visual-linguistic data [9-11]. Overall, multimodal aspect-level sentiment analysis has achieved better results. However, its application in depression detection is still relatively limited.

Therefore, this paper concentrates on utilizing multimodal aspect-based sentiment analysis for the detection of depression. It uses the model proposed by YU et al. for depression detection, known as the Target-Oriented Multimodal Bidirectional Encoder Representation from Transformers (TomBERT) [12]. This paper's experimental results show that multimodal aspect-level sentiment analysis can better detect depressive tendencies, compared with other methods.

2 Methodology

2.1 Dataset

The data source used in this paper is a public dataset published by Gui's team in 2019 [13]. The dataset consists of two main components: images and JSON data. The meanings of the keys and values in the JSON data are presented in table 1.

Table 1. Keys and values in the JSON data.

Key	Value
ID	Data ID
Id_str	The corresponding image ID
Text	The text of the tweet
Source	The device name
User	Username and Twitter ID

In this dataset, some tweets do not have corresponding images. To facilitate research on multimodal sentiment analysis, this paper removes tweets without images from the dataset. In addition, data deduplication is performed, and unnecessary labels are deleted. Table 2 shows the processed JSON data of this dataset.

Table 2. Labels after dataset cleaning.

Key	Value
Index	Data ID
Label	Label of depressive tendencies
Image_ id	The corresponding image ID
Aspect	aspect word
Text	The text of the tweet

The original dataset lacks annotations for aspect words and depressive tendencies of tweets. To conduct aspect-level sentiment analysis for depression tendency, this paper annotates aspect words and depressive tendencies of the tweets in the dataset.

Table 3 randomly shows two annotated data samples. Here, ‘Aspect’ denotes aspect words, and ‘T’ marks their positions within the complete text of the tweet. Aspect words are fundamental elements in sentiment analysis, as they help determine the specific object in the text that the sentiment is being expressed towards. By employing methods to annotate aspect words, further research can be conducted on aspect-level sentiment analysis, thereby gaining a deeper understanding of people's emotional tendencies towards different objects or attributes.

Table 3. Two annotated data entries.

Index	Label	Image_Id	Aspect	text
6259	0	502536185252937728.jpg	i	T drew this for you! T hope you get to 2 MILLION! @TheMattEspinosa # MattTo2Mill
5049	1	817168866413514752.jpg	kids	RT @imgur: When only one of your T can read.

'Label' indicates the annotations of depressive tendencies. Table 4 demonstrates the sample composition of the processed dataset. There are 4401 samples for non-depressive tendencies and 5262 samples for depressive tendencies.

Table 4. Composition of samples in the dataset.

Label	Number
0（non-depressive tendency）	4401
1（depressive tendency）	5262

2.2 Model for multimodal aspect-based sentiment analysis

In this paper, the multimodal aspect-level sentiment analysis model TomBERT is applied to the social platform dataset. Additionally, this paper conducts comparative experiments using both Multimodal Bidirectional Encoder Representation from Transformers (mBERT) model proposed by YU et al and the aspect-level sentiment analysis-based BERT model to study the effectiveness of the multimodal aspect-level approach.

TomBERT model. TomBERT is a sophisticated model designed for multimodal aspect-level sentiment analysis, whose structure is shown in Fig. 2. For textual data, it first utilizes BERT to obtain text word embeddings, which are enriched with semantic and contextual information of the text. Subsequently, a Sentence Encoder is employed to extract aspect-specific word features from the text. Concurrently, a Target Encoder is used to process the opinion target words. For image data, ResNet-152 is applied to extract features. With its powerful feature representation capabilities and depth, ResNet-152 is able to capture key information and details in the images, providing strong visual support for sentiment analysis. Following that, cross-attention is utilized, with the target text as a mask, to perform attention on the images, resulting in target-image attention embeddings. This step effectively integrates textual and visual information, enabling the model to comprehensively understand and analyze sentiment tendencies. Finally, the target-image hybrid features are concatenated with the textual aspect-specific word features, and multiple layers of self-attention are stacked to further capture and strengthen the correlations and dependencies between these features. The predicted output is then obtained through a pooling layer, a linear layer, and a softmax layer.

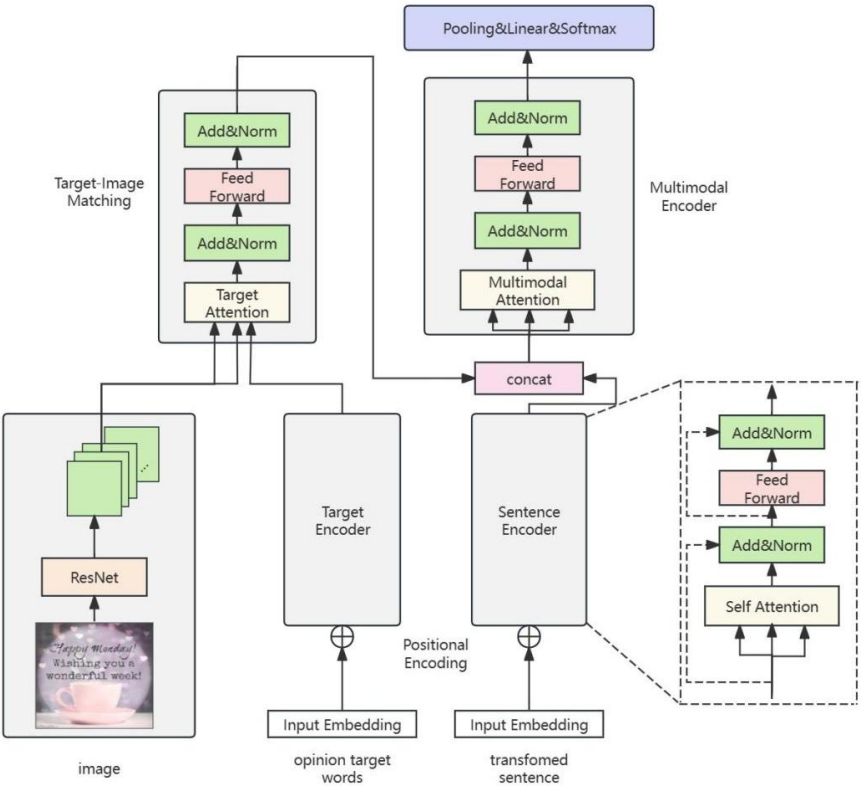

Fig. 2. The structure of TomBERT (Photo credit: Original).

TomBERT combines data from both text and image modalities and performs sentiment analysis at the aspect level. The model possesses fine-grained language understanding capabilities and enhances the accuracy and robustness of sentiment analysis by introducing data from the image modality.

mBERT model. mBERT is a multimodal non-aspect-level model, whose structure is depicted in Fig. 3 Compared to the TomBERT model, the mBERT model does not incorporate aspect-level inputs. In the mBERT model, for image modality data, it utilizes ResNet to extract key features and details from the images. It also employs a Sentence Encoder to process textual modality data, extracting overall semantic and emotional information from the text. Afterwards, the text features and image features are concatenated for fusion. The fused features are then processed by a Multimodal Encoder for further encoding and transformation to better suit subsequent classification tasks. Finally, a pooling layer, a linear layer, and a softmax layer are utilized to obtain the predicted output.

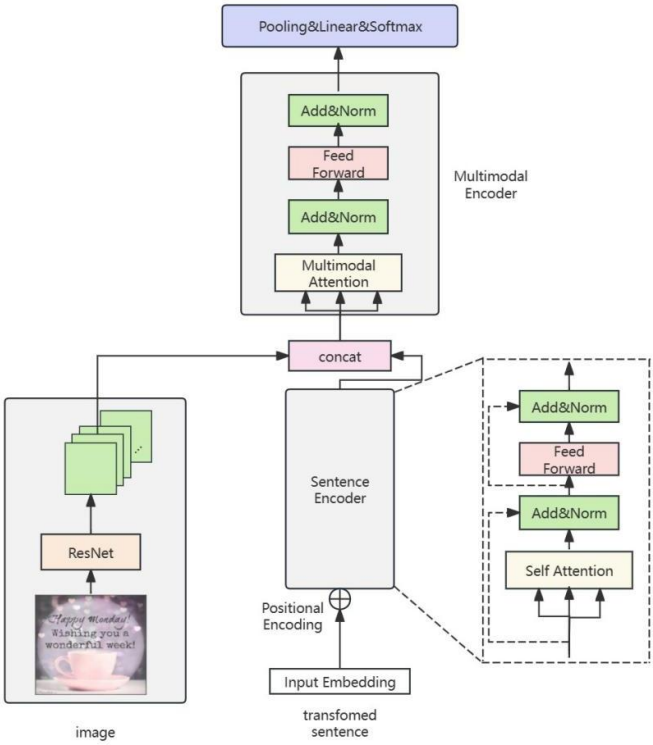

Fig. 3. The structure of mBERT (Photo credit: Original).

With the help of the ResNet branch, mBERT can simultaneously process data from both text and image modalities. This approach enables dual-modal sentiment analysis. Multimodal methods allow it to understand richer features in tweets, but they utilize a coarser granularity, which may potentially affect the detection effectiveness.

BERT model. BERT, introduced by Google researchers in 2018, is a pre-trained language model built on Transformers [14]. The BERT model has significantly enhanced performance in various natural language processing (NLP) tasks, including named entity recognition, text classification, sequence tagging, sentiment analysis, and more. As shown in Fig. 4, BERT makes use of the Transformer's encoding component.

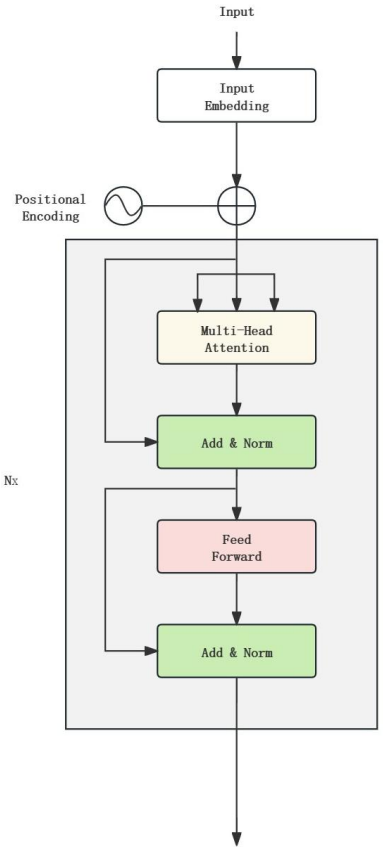

Fig. 4. The encoding part of the Transformer used by BERT (Photo credit: Original).

The primary advantage of the BERT model is its bidirectional encoding capability, which allows it to consider both preceding and succeeding contextual information during training. Compared to traditional language models, BERT leverages pre-training on vast amounts of unlabeled text data to acquire rich linguistic knowledge, which is then applied to various specific NLP tasks through fine-tuning.

The BERT model can be used to perform aspect-level sentiment analysis by taking both text and aspect words as input. The aspect words originate from the words in the sentence, and after extracting the aspect words, the position of that word is replaced with a special token 'T'. After the BERT model processes the input, a fully connected layer and a Tanh activation function are applied in the downstream stage to produce the output, achieving the output sentiments. The structure of the BERT model is shown in. Fig. 5.

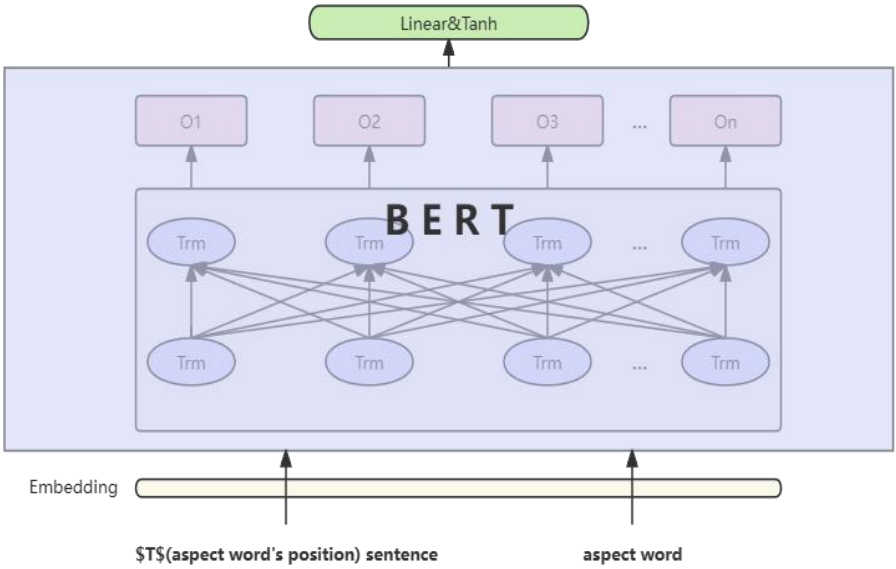

Fig. 5. The structure of BERT (Photo credit: Original).

3 Results and Discussion

3.1 Training

In this paper, the dataset introduced in Section 2.1 is used for training. During the training process, the data is allocated into three sets: training, validation, and test, following a 7:1:2 ratio.

The training process was conducted within the PyTorch deep learning framework. The PyTorch version is 2.1.2 and the CUDA version is 12.1. The experiments were conducted on a Linux-based system equipped with a NVIDIA Tesla T4 GPU, and 16GB of the RAM. The model was trained using a batch size of 32. A batch size of 32 allows for efficient use of GPU memory while also providing enough samples for the model to learn from during each iteration. The training epochs was set to 8. The initial learning rate was set to 5e-5. To prevent overfitting, a dropout rate of 0.1 was applied. The number of attention heads was set to 16.

3.2 Evaluation and Results

After training, this paper uses accuracy and F-score as evaluation metrics to assess the detection results. Specifically, accuracy represents the ratio of correctly predicted tweets to the total number of tweets. The higher this ratio is, the more accurate the prediction is. The F-score can comprehensively measure the precision of the prediction results. The higher this value is, the more accurate the prediction is. The formulas for accuracy and F-score are shown in Equations 1 and 2.

$$Accuracy = \frac{TP}{TP+FP} \tag{1}$$

$$\mathrm{F-score} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \quad (2)$$

Where TP represents the count of actual depression tendencies, FP indicates the number of incorrectly identified depression tendencies, and FN denotes the number of misidentified non-depression tendencies. To investigate the detection performance of different methods on social media datasets, this paper conducted experiments using three methods: the TomBERT model introduced in Section 2.2.1, the mBERT model described in Section 2.2.2, and the BERT model described in Section 2.2.3. TomBERT is an aspect-level and multimodal model, mBERT is a non-aspect-level, multimodal model, while BERT is an aspect-level, unimodal model. The evaluation metrics mentioned above were utilized to evaluate the prediction outcomes on the test set. The experimental findings are displayed in table 5.

Table 5. Experimental results.

Model Name	Accuracy	F-score
TomBERT（aspect-level, multimodal）	89.67%	89.58%
mBERT（non-aspect-level, multimodal）	88.43%	88.33%
BERT（aspect-level, unimodal）	87.18%	87.05%

As shown in table 5, the TomBERT model, employing an aspect-level multimodal approach, achieved the best results. Compared to the mBERT model which is not an aspect-level method, it achieved a 1.24% increase in accuracy and a 1.25% increase in F-score. This demonstrates the superiority of aspect-level methods over non-aspect-level methods in detecting depressive tendencies. Compared to the unimodal BERT model, it demonstrated a 2.49% improvement in accuracy and a 2.53% enhancement in the F-score. This demonstrates the superiority of multimodal methods over unimodal methods in detecting depressive tendencies. The experimental results demonstrate that the aspect-level multimodal sentiment analysis approach achieved the best depression tendency detection performance on the social media dataset.

4 Conclusion

Depression is a psychological illness that poses a serious threat to both physical and mental health. Traditional diagnostic processes are complex and time-consuming, relying on self-evaluations and expert assessments, which complicates early detection and intervention. Recently, the detection of depressive tendencies using social media data has shown potential, particularly through sentiment analysis. This article explores the application of aspect-level multimodal sentiment analysis in depression detection, utilizing the TomBERT model to analyze social media data and comparing it with other models. Experimental results demonstrate that aspect-level multimodal sentiment analysis outperforms other methods, providing a deeper understanding of users' emotional tendencies and enabling more accurate identification of depressive tendencies.

However, this study has limitations. First, the dataset's aspect words were manually classified by the authors. Future research should focus on developing better tools for aspect-word extraction and sentiment analysis. Second, depressive tendencies in this study were labeled as either present or absent. In reality, depressive tendencies vary in severity, and future research should explore multi-class classification to achieve more nuanced identification, such as "mild," "moderate," and "severe" depression.

References

1. Hasib, K. M., Islam, M. R., Sakib, S., Akbar, M. A., Razzak, I., Alam, M. S.: Depression Detection from Social Networks Data Based on Machine Learning and Deep Learning Techniques: An Interrogative Survey. IEEE Transactions on Computational Social Systems 10(4), 1568–1586 (2023).
2. Wang, X., Chen, S., Li, T., Li, W., Zhou, Y., Zheng, J., Chen, Q., Yan, J., Tang, B.: Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis. JMIR Medical Informatics 8(7), e17958 (2020).
3. Zhu, X., Huang, Y., Wang, X., Wang, R.: Emotion recognition based on brain-like multimodal hierarchical perception. Multimedia Tools and Applications, 1–19 (2023).
4. Zhao, H., Yang, M., Bai, X., Liu, H.: A Survey on Multimodal Aspect-Based Sentiment Analysis. IEEE Access 12, 12039–12052 (2024).
5. Makiuchi, M. R., Warnita, T., Uto, K., Shinoda, K.: Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. In Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge (2019).
6. Chiu, C., Lane, H.-Y., Koh, J.-L., Chen, A. M.: Multimodal Depression Detection on Instagram Considering Time Interval of Posts. Journal of Intelligent Information Systems 56 (2021).
7. Bengio, Y., Yu, X., Shuwen, Z.: Aspect-Based Sentiment Analysis Model of Multimodal Collaborative Contrastive Learning. Data Analysis and Knowledge Discovery, 1–16 (2023).
8. Yu, J., Chen, K., Xia, R.: Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing 14(3), 1966–1978 (2023).
9. Li, L., Li, P.: Aspect-Level Multimodal Sentiment Analysis Based on Interaction Graph Neural Network. Applied Research of Computers 40(12), 3683–3689 (2023).
10. Zhou, R., Zhu, H. Z., Guo, W. Y., Yu, S. L., Zhang, Y.: A Unified Framework for Multimodal Aspect-Term Extraction and Aspect-Level Sentiment Classification. Journal of Computational Research and Development 60(12), 2877–2889 (2023). [doi: 10.7544/issn1000-1239.202220441]
11. Ling, Y., Yu, J., Xia, R.: Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2149–2159 (2022).

12. Yu, J., Jiang, J.: Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In International Joint Conference on Artificial Intelligence (2019).
13. Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., Chen, Z.: Cooperative Multimodal Approach to Depression Detection in Twitter. In Proceedings of the AAAI Conference on Artificial Intelligence 33(01), 110–117 (2019).
14. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018).