ITC 2/54
Information Technology and Control
Vol. 54 / No. 2/ 2025
pp. 451-470
DOI 10.5755/j01.itc.54.2.39960

MEA-IFE: An Improved Multi-modal Fusion Framework Based on DCNN-BERT-BiLSTM and Its Application in Sentiment Analysis

Received 2024/12/27 | Accepted after revision 2025/03/13

HOW TO CITE: Ye, H., Xiao, X. (2025). MEA-IFE: An Improved Multi-modal Fusion Framework Based on DCNN-BERT-BiLSTM and Its Application in Sentiment Analysis. *Information Technology and Control*, 54(2), 451-470. https://doi.org/10.5755/j01.itc.54.2.39960

MEA-IFE: An Improved Multi-modal Fusion Framework Based on DCNN-BERT-BiLSTM and Its Application in Sentiment Analysis

Hongfei Ye
School of Intelligent Science and Technology; Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences; Hangzhou, China; 310024; phone: 15068508642; e-mail: yehongfei23@mails.ucas.ac.cn

Xiaochen Xiao
TD School; University of Technology Sydney (UTS); Sydney, Australia; 2007; phone: 13715280886; e-mail: x971851309@gmail.com

Corresponding author: yehongfei23@mails.ucas.ac.cn

In the real world, emotional data often comes from multiple heterogeneous sources, making it difficult for unimodal approaches to capture emotional information fully. Existing sentiment analysis models struggle with accuracy when handling complex emotional expressions. Accordingly, this paper proposes a multi-modal sentiment analysis framework, MEA-IFE, which is characterized by effective feature extraction and high predictive accuracy. To mitigate potential information loss and expression limitations in BERT-BiLSTM during text feature extraction, MEA-IFE introduces a parallel structure of SK-Net and BiLSTM, enhancing the ability to extract multi-dimensional text features. Additionally, it integrates the ECA mechanism to improve the capture of essential information in text. For image-related challenges, MEA-IFE incorporates Vision Transformer better to capture both global and detailed features of images, combining CNN and Transformer architectures. During the feature fusion phase, MEA-IFE employs a multi-head attention mechanism to dynamically integrate text and image features, exploring the interactive potential be-

tween different modalities. Experiments performed using the Kaggle text dataset and the FER2013 image dataset demonstrate an impressive accuracy of up to 98.00%, validating its effectiveness. When compared with models like AM-MF, AMSAER, HAN-CA-SA, and TBGAV, MEA-IFE shows outstanding performance across accuracy, precision, recall, and F1 score, with respective improvements of 0.40%, 0.20%, 0.75%, and 0.52%. The model also excels in the AUC metric, further confirming its advantages. The proposed MEA-IFE model possesses high predictive accuracy and strong feature integration capabilities, meeting the precision demands of complex multi-modal sentiment tasks.

KEYWORDS: Feature Extraction, Multi-head Attention, Sentiment Analysis, Multimodal Fusion

1. Introduction

Sentiment analysis is a critical domain within NLP that focuses on the automatic detection of emotional tendencies in data. It involves the extraction, analysis, and mining of subjective data that is infused with emotional nuances [5]. Owing to the rapid expansion of social media and online reviews, sentiment analysis is being applied more and more extensively across various sectors, including business intelligence, market research, and public safety [15]. However, emotional data in the real world often originates from multiple heterogeneous sources, including text, images, and voice. As a result, unimodal analysis methods struggle to comprehensively capture emotional information, making the effective integration of multi-modal information a key challenge [11]. Additionally, existing sentiment analysis models still face issues of low accuracy when dealing with complex emotional expressions, which do not fulfil the requirements of real-world applications [4].

Early research on sentiment analysis primarily focused on text datasets. Tang et al. [6] proposed an LSTM-based model that demonstrated high accuracy on Twitter datasets, particularly when considering the semantic relevance between target words and their contextual counterparts. Nevertheless, this approach had certain limitations in training time and was highly dependent on the datasets used. Salem et al. [12] introduced Slim LSTM, which reduced the number of parameters to accelerate training and lower computational costs. This model achieved good results on the GOP debate Twitter dataset but still faced challenges in handling imbalanced datasets and complex emotional expressions. Wang et al. [19] introduced a method which combines text filtering networks with an enhanced BERT model, resulting in a significant improvement in classification accuracy on the sentiment analysis dataset from the AI Challenger. However, issues related to redundancy and noise in long texts persisted, necessitating further improvements in text filtering accuracy and the effective processing of extended text information. Wu et al. [1] introduced the Quasi-Attention Context-Guided BERT model, which significantly improved sentiment analysis performance by integrating contextual information, demonstrating excellent results on the SentiHood and SemEval-2014 datasets. Nevertheless, text sentiment analysis is easily influenced by context and semantic factors, making it challenging to reflect human emotions accurately.

Image sentiment analysis captures non-verbal emotional signals, such as facial expressions and postures, providing a more intuitive and objective reflection of emotional states. This research field, which centers on recognizing and analyzing sentiment in images, offers numerous application possibilities but encounters considerable challenges due to the abstract characteristics of visual content. Cao et al. [8] introduced a method for image sentiment classification utilizing Adaboost combined with BP neural networks, which achieved good classification results compared to traditional BP neural network methods, particularly in enhancing the efficiency and accuracy of semantic classification of emotional images. Xu et al. [7] presented an innovative framework for visual sentiment prediction based on DCNN, which effectively performed sentiment analysis through transfer learning from a DCNN that had been pre-trained on a large dataset, significantly improving sentiment prediction performance. However, further exploration is needed to optimize the model for better robustness against noisy data. You et al. [21] proposed an image sentiment analysis model based on CNN, which employed progressive training and domain transfer strategies

to mitigate the impact of noisy data on model performance. This model demonstrated excellent results on the Flickr and Twitter datasets but still suffered from limitations in effectively extracting global features. Zhu et al. [23] developed a cohesive CNN-RNN model that extracted features at various levels through CNN in a multi-task learning framework and performed feature fusion using a Bi-RNN, thereby improving the performance of sentiment recognition.

The studies mentioned above focus solely on single-modal data sources, overlooking the interactions between multiple modalities. Compared to unimodal approaches, multi-modal sentiment analysis methods exhibit higher performance. Wang et al. [18] introduced the CCLA sentiment classification model, which effectively captures local and long-range semantic and emotional information in text by combining CNN, LSTM and Attention. Tang et al. [17] introduced a model named CTFN, which captures bidirectional interactions between different modalities using a coupled learning approach. This model integrates a cyclic consistency constraint to enhance performance, allowing it to retain only the encoder part of the Transformer, thereby reducing complexity. Zhang et al. [22] presented the TBGAV model, which employs TinyBERT and BiGRU-Attention to extract and enhance text features, utilizes VGG19 to extract visual features, and applies a bilinear fusion method to improve multi-modal sentiment classification performance. Similarly, Xie et al. [20] introduced the AM-MF model, which combines RoBERTa, ResNet, and Vision Transformer for multi-modal feature extraction. This model builds a CMA interaction component that enables the integration of features between various modalities and employs a soft attention strategy for deeper integration of both internal and cross-modal information, significantly improving sentiment classification accuracy. Nevertheless, these approaches still encounter challenges, such as being overly simplistic and ineffective in feature extraction, indicating the need for more effective models to meet the demands of complex sentiment analysis tasks.

Analyzing sentiment across multiple modalities involves combining features from various sub-modalities to capture emotional information more comprehensively. Dushyant et al. [25] utilized an IIM module to learn interactions between different modalities, enhancing model performance with cross-modal features and introducing a CAM module to capture contextual details, thereby enhancing the accuracy of sentiment and emotion analysis. Ameer et al. [2] introduced a late fusion-based multi-modal multi-task learning framework, integrating three late fusion models (TFN, LMF, and LF-DNN) into this multi-task learning structure, resulting in significant improvements in multi-modal task performance. Zhang et al. [9] employed a transformer-based unimodal encoder to extract features and introduced a soft modal attention mechanism during the fusion stage of the features. This mechanism dynamically assigns different importance weights to each emotion label, capturing the dependency relationships between various modalities and emotion labels. They modeled the bidirectional interactions between modalities using a coupled learning approach within the encoder. Zhao et al. [24] designed a unified multi-modal prompting method and employed probabilistic fusion techniques to aggregate predictions based on different modal prompts, effectively mitigating the impact of feature discrepancies between modalities. Despite certain advances in current multi-modal fusion methods, there remain shortcomings in fusion flexibility and cross-modal contextual modeling. Therefore, an efficient and flexible fusion mechanism must be designed to enhance the representational capability of the model.

In conclusion, this study presents a Multi-modal Emotion Analyzer with an Integrated Feature Enhancement model, which is characterized by effective feature extraction capabilities and high predictive accuracy. Specifically, to address the potential issues of information loss or limited expressive power in text feature extraction by BERT-BiLSTM, the paper introduces the SK-Net to be integrated into parallel with Bi-LSTM to improve the multi-scale spatial feature extraction abilities of the model. The parallel integration of SK-Net and Bi-LSTM can extract richer and more comprehensive feature representations from multiple perspectives. Subsequently, based on BERT-SKNet-BiLSTM, the Efficient Channel Attention is introduced to improve the text feature extraction module's ability to capture important information. To address the issue of insufficient global feature capture in the DCNN AlexNet model for image information extraction, this paper introduces the Vision Transformer module to enhance the global representation in images. In the feature fusion part, multi-head attention is combined to adaptively merge cross-modal text and image features, fully ex-

ploring the interaction between different modalities, which significantly enhances the accuracy of emotion and sentiment analysis. This study evaluates the proposed model against four other models, AM-MF, AM-SAER [3], HAN-CA-SA [16], and TBGAV, to demonstrate its effectiveness and superiority.

The contributions of this research are outlined as follows:

1 A BERT-SKNet-BiLSTM-ECA module is proposed for text feature extraction. This module integrates BERT, SK-Net, BiLSTM and ECA to derive richer and more comprehensive text features, addressing potential information loss and limited expressiveness in BERT's feature extraction.

2 For image feature extraction, a DCNN+Vision Transformer model is proposed, which can enhance the ability to capture global representations in images, compensating for the insufficiency of DCNN models in capturing global features in image information extraction.

3 The MEA-IFE model can enhance the prediction accuracy of multi-modal sentiment evaluation. Through the combination of multi-head attention to adaptively fuse cross-modal text and image features, it explores the interaction relationships between different modalities, effectively improving feature integration and overall model performance.

This paper is organized as follows: Section 2 presents the proposed models and methods, including the text and image feature extraction modules fusion mechanism, along with a comprehensive overview of the components within each module. Section 3 is the experimental part, which introduces the simulation environment, evaluation indicators, and datasets used in the experiments and proves the efficacy and advantages of the proposed model through ablation experiments and comparative experiments. Section 4 will discuss the advantages, limitations, application prospects, and future work of the proposed model.

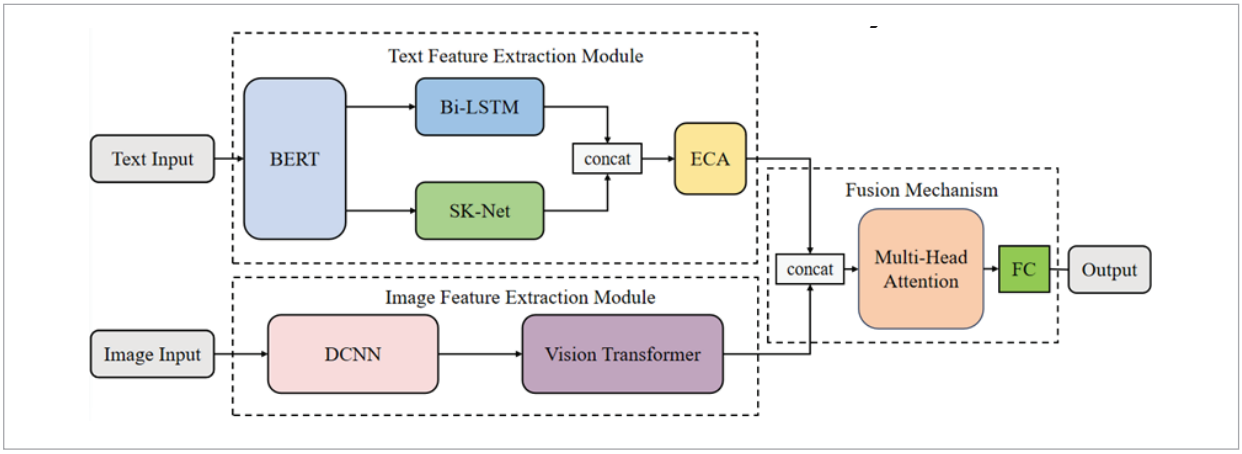
2. Models and Methods

The framework of the MEA-IFE multi-modal sentiment analysis model proposed in this study, illustrated in Figure 1, comprises the text and image feature extraction module, along with the multi-modal fusion part.

2.1 Text Feature Extraction Module

After preprocessing, the original text data is input into BERT, which can comprehend the contextual nuances within the text, capture the semantic relationships between words, phrases, and sentences, and generate a high-dimensional feature representation. Subsequently, the features generated by BERT will be sent into two parallel integrated modules: Bi-LSTM and SK-Net. The Bi-LSTM module can handle the contextual details within the text, capturing the global information in the content and further enhancing the model's understanding of sequential data. The SK-

Figure 1
MEA-IFE Multi-modal Sentiment Analysis Framework Diagram.



Net module can adaptively extract features at different scales, effectively enhancing the model's capacity to identify local characteristics. The features extracted by the Bi-LSTM and SK-Net modules will then be passed into the ECA module for feature fusion. ECA can efficiently identify and enhance important channel features, and this lightweight architecture can decrease the model's complexity while maintaining good performance.

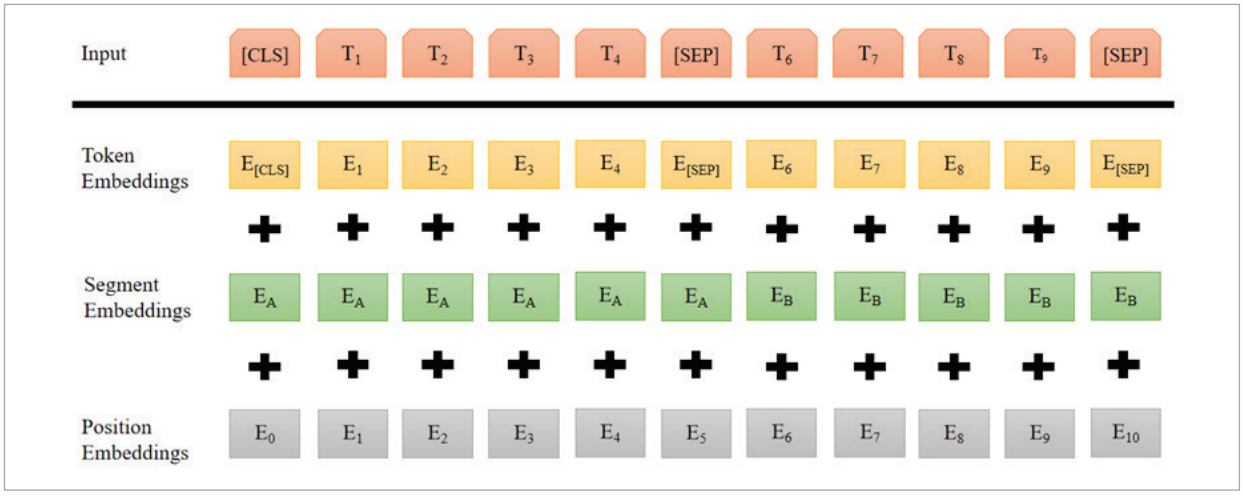
2.1.1 BERT

BERT is a pre-trained LM proposed in 2018, built on the Transformer architecture. In contrast to conventional unidirectional language models like Word2Vec and ELMo, BERT uses a bidirectional training method, which means that when processing each word, it considers the words to the left and right simultaneously. It allows the model to gain a deeper insight into the context, capture bidirectional relationships between words, and thus generate richer context-related word vectors. BERT utilizes the encoder component of the Transformer, relying on the self-attention mechanism to process sequential data, effectively capturing long-distance dependencies. BERT's pre-training consists of two primary tasks: MLM and NSP. In MLM, certain words in the input sentences are randomly masked, and the model is tasked with predicting them; NSP asks the model to predict whether two sentences are consecutive. BERT has attained breakthrough performance across a range of NLP tasks, becoming a new standard for many tasks, and has driven the research direction of pre-trained language models, having a profound impact on the entire field.

BERT's architecture primarily features an embedding layer along with 12 stacked Transformer encoders. After tokenization and preprocessing, including the addition of [CLS] and [SEP] tokens, the text is fed into the embedding layer. This layer comprises three main parts: Token Embedding, Segment Embedding, and Position Embedding. Token Embedding represents the tokenized words as vectors in a high-dimensional space, while Segment Embedding helps differentiate between different sentences. Position Embedding is introduced to give the model insights into each token's position within the sequence, allowing it to grasp the relationships between words. The integration of these three types of embeddings equips BERT to handle complex language structures and contextual relationships.

The text data represented by the embedding layer is transmitted to the cascaded Transformer encoders, where each encoder layer includes self-attention and FNN, enabling the model to identify intricate dependencies within the input sequence. In the BERT model, self-attention is one of the core components, capable of capturing long-distance dependencies between different tokens and constructing richer global semantic feature representations. This mechanism enables the model to take into account other tokens throughout the entire input sequence while processing each token, thus achieving a more accurate seman-

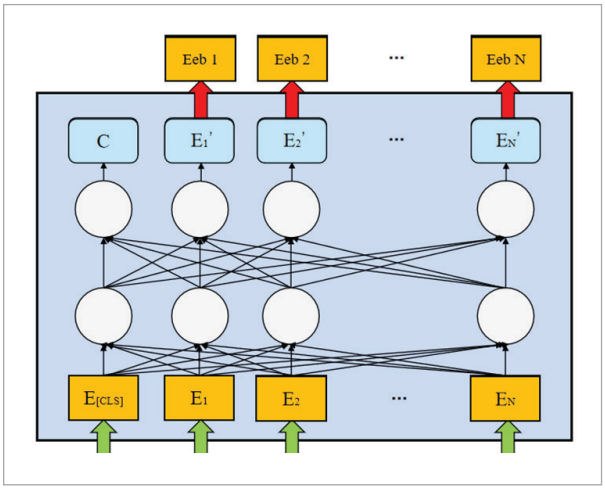
Figure 2
Structural Representation of the BERT Embedding Layer.



tic understanding. The encoder module also includes normalization layers, which can help stabilize the training process and accelerate convergence. Normalization layers typically use Layer Normalization, which can reduce the issue of internal covariate shift, thereby enhancing the stability of the model throughout the training process. The FNN further nonlinearly transforms the features processed by self-attention and normalization to enhance the semantic expression ability of the features. Additionally, residual modules are introduced in the encoder module, which assists in addressing the issues of gradient vanishing or explosion in deep networks, ensuring that information can be effectively transmitted within the network. By repeating the above process in multiple Transformer layers, the model can gradually refine and accumulate high-level semantic features of the text, ultimately outputting an in-depth understanding of the text.

Figure 3

Framework Diagram of BERT-based Feature Extraction.



2.1.2 Bi-LSTM

Bi-LSTM is a particular variant of RNN architecture that integrates LSTM units with bidirectional processing mechanisms. By adding bidirectional operations to the foundation of LSTM, Bi-LSTM is capable of simultaneously handling both preceding and subsequent information in sequence data, thereby enhancing its ability to capture dependencies within the sequence. LSTM consists of three main components: the input gate, which regulates new information retained in the cell state; the forget gate, which determines information to discard; and the output gate, which dictates how the next hidden state uses the current cell state. The coordinated operation of these three gates allows LSTM to retain and leverage long-term dependencies within the sequence, mitigating the vanishing and exploding gradient issues common in RNNs. The relationships mentioned above are further expressed through formulas as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (2)$$

$$\widetilde{C_t} = \tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (3)$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t} \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (5)$$

$$h_t = o_t \times \tanh(C_t), \quad (6)$$

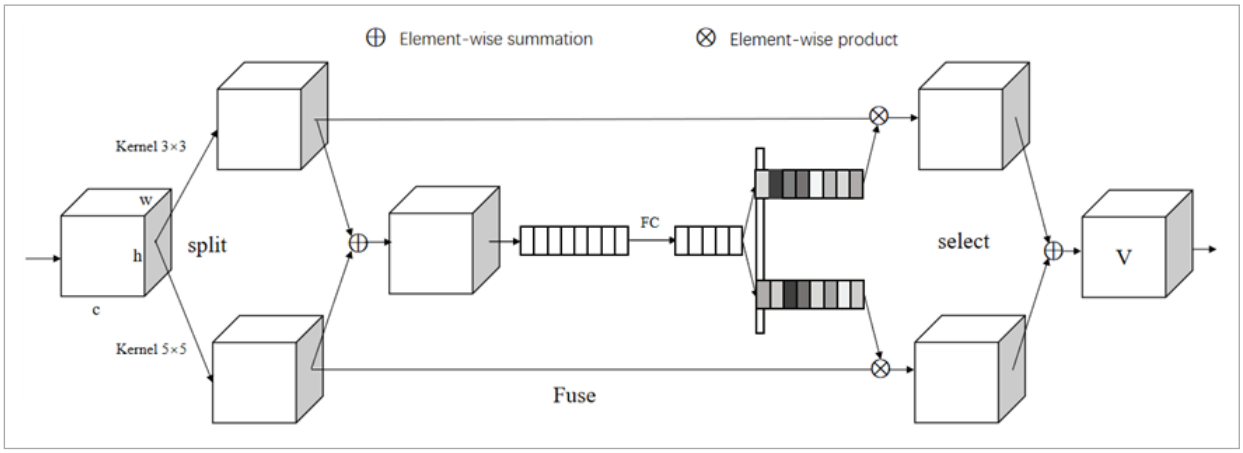
where represents the current time step, denotes the hidden state of the previous time step, indicates the cell state of the prior time step, and refers to the weight matrix and bias term, signifies the sigmoid function and represents the hyperbolic tangent function.

The Bi-LSTM module adds bidirectional processing capabilities to the foundation of the LSTM. At each step, the model is capable of processing not only the forward information associated with the current time step but also the information from preceding time steps. It means that for each element in the sequence, the Bi-LSTM will generate two hidden states. This bidirectional processing mechanism enables the model to consider contextual information simultaneously, thus offering a deeper understanding of sequence data.

2.1.3 SK-Net Mechanism

SK-Net (Selective Kernel Network) is a CNN architecture designed to adaptively select features across multiple scales, thereby improving the model's capacity to capture information at diverse scales. Traditional convolutional neural networks use fixed-size convolutional kernels and cannot dynamically modify the size of the receptive field. SK-Net achieves adaptive receptive fields by dynamically adjusting the size of the convolutional kernels, thereby better-capturing features at different scales. The core idea is to use a set of convolutional kernels of different sizes that can

Figure 4
SK-Net Architecture Diagram.



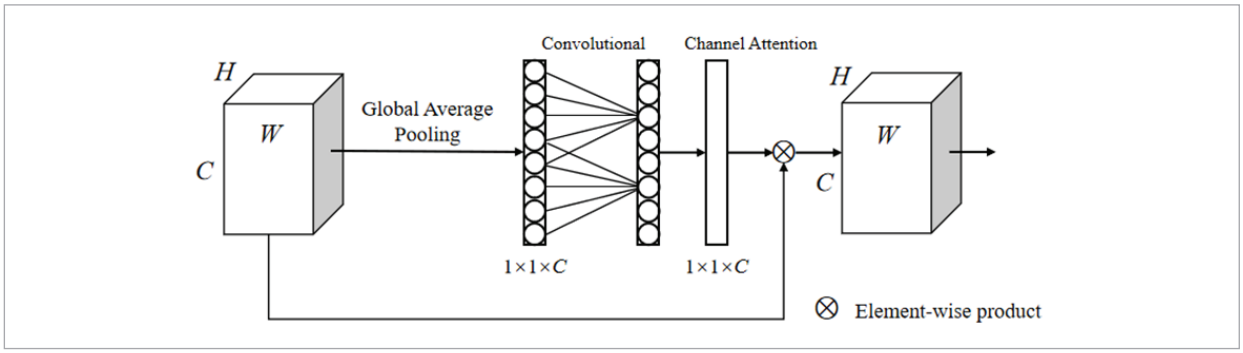
extract local features of the data in parallel. Then, the model subsequently incorporates a channel attention mechanism that adaptively allocates varying weights to the features extracted by different convolutional kernels. This selective mechanism allows SK-Net to flexibly handle features of different scales without the need for manual design or preset. The main architecture is shown in Figure 4.

SK-Net is commonly used in computer vision tasks for extracting image features, but its adaptive feature extraction capability can also be applied to text feature extraction. For text data, these convolutional kernels of different sizes can capture features at different granularities within the text, greatly enhancing the extraction of local textual features. The integration of the SK-Net module in parallel with the Bi-LSTM module can fully extract features at different granularities in the text, taking into account both local and global features, which achieves more comprehensive feature extraction and sequence modeling.

2.1.4 ECA Mechanism

ECA is a lightweight channel attention mechanism that effectively improves upon traditional channel attention mechanisms. The core advantage of ECA lies in its significant reduction in model parameters and overall computational complexity through the simplification of the computation process, thereby enhancing the operational efficiency of the model. Traditional channel attention mechanisms, such as SENet and CBAM, typically rely on two fully connected layers to calculate the channel attention weights. However, ECA simplifies the structure and reduces the model's complexity through the substitution of fully connected layers with one-dimensional convolutional layers.

Figure 5
Flowchart of the ECA Mechanism.



After the text data has been processed through the Bi-LSTM module and SK-Net for feature extraction, these features are sent to the ECA module. The ECA module functions to automatically recognize and enhance important channel features while suppressing less critical information. This mechanism enables the model to prioritize essential features, facilitating the effective integration of global features obtained from Bi-LSTM and local features derived from SK-Net, thus contributing to an enhancement in the model's accuracy and further improving its generalization capability and practical applicability.

2.2 Image Feature Extraction Module

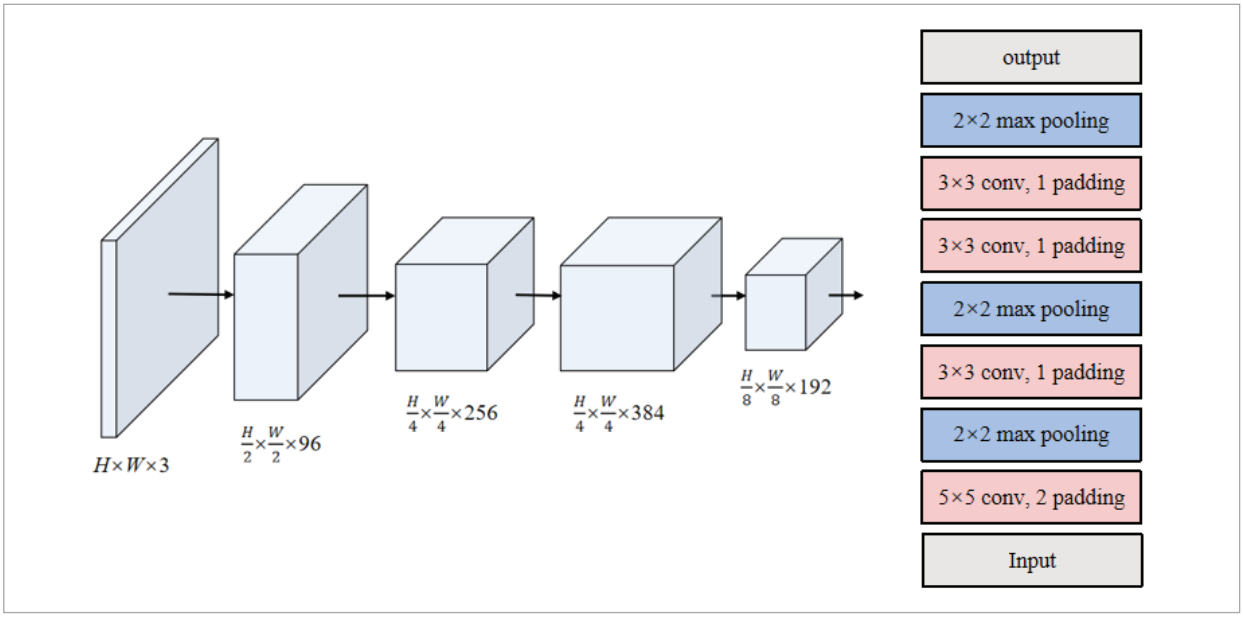
The image feature extraction module is composed of DCNN and Vision Transformer. The DCNN part utilizes the AlexNet architecture, which extracts local features from images via a sequence of convolutional layers, activation functions, and pooling layers. To address the issue of insufficient global feature capture in DCNN-based image feature extraction, this paper introduces a Vision Transformer module that employs self-attention techniques to identify global relationships and long-range dependencies within images. By integrating CNNs with Transformers, the representational capacity of both global and local features in images is enhanced.

2.2.1 DCNN

DCNN is a deep learning architecture that effectively learns high-level features of images and is widely used in computer vision because of its powerful feature extraction abilities. Its fundamental characteristic lies in the arrangement of various convolutional layers, pooling layers, and FC layers. Convolutional layers extract local features from images while pooling layers decrease the dimensionality of these features and enhance their invariance. FC layers integrate the features obtained from the previous layers. In this context, the AlexNet architecture [13] is employed as a feature extractor to capture potential emotional features within images. To further explore the emotion-related features in image data, we have adjusted some parameters of the AlexNet model and removed the FC layers and classification parts, retaining only the feature extraction part of the network to concentrate on extracting semantic information related to the emotional features contained in the images. The main framework is shown in Figure 6.

AlexNet is a DCNN architecture introduced by Alex Krizhevsky et al. in 2012 [13]. One of its core features is the use of the ReLU activation function, as shown in Equation (7). This choice can markedly enhance the model's training speed. Compared with tradi-

Figure 6
AlexNet Feature Extraction Framework [13].



tional Sigmoid or Tanh activation functions, ReLU reduces computational complexity and alleviates the vanishing gradient problem, making it more efficient for processing deep networks. In addition, AlexNet also introduced the Dropout regularization technique, which serves as a straightforward yet effective approach that forces the network to learn more robust feature representations, thereby enhancing the model's generalization capability and reducing the risk of overfitting.

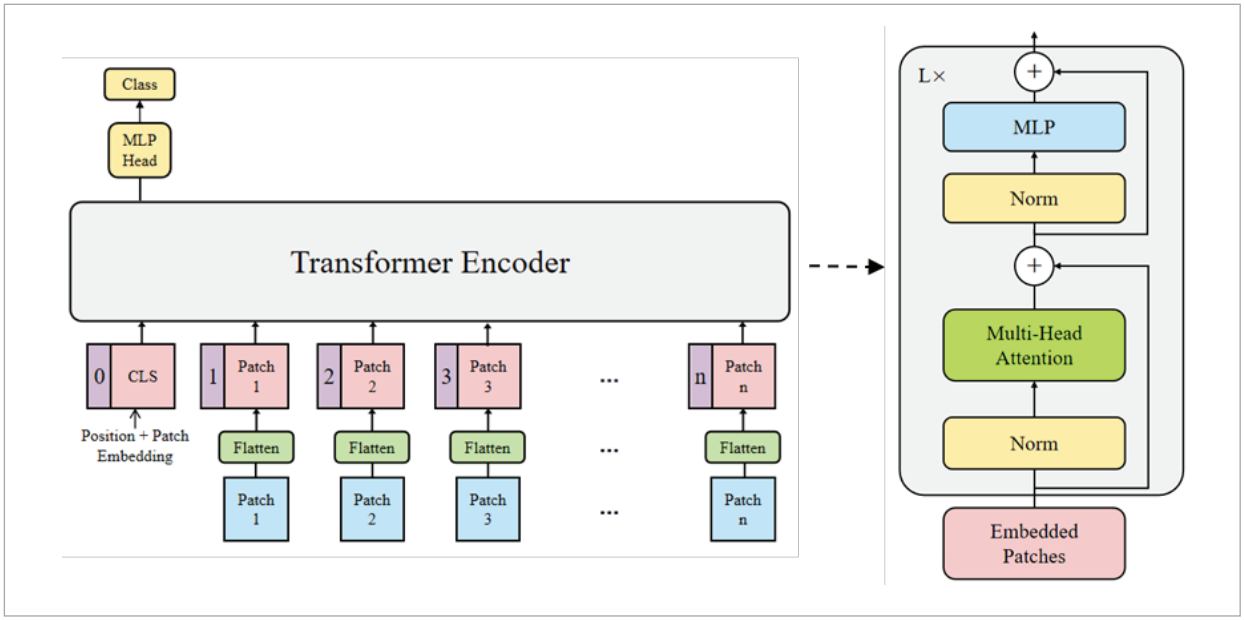
$$\mathrm{ReLU}(x) = \begin{cases} x \text{ if } x > 0 \\ 0 \text{ if } x \le 0 \end{cases} \tag{7}$$

2.2.2 Vision Transformer

Vision Transformer (ViT) is an innovative deep-learning architecture that successfully extends the advanced capabilities of the Transformer framework to the domain of image processing. Its core innovation lies in transforming traditional image processing tasks into sequence modeling problems, allowing the model to capture both global and local features of images using the Transformer.

The image features obtained from DCNN are passed to the ViT module. ViT first divides the input image into multiple patches, which can be considered as local features of the image. Each patch is transformed into a one-dimensional vector and then embedded via a linear layer to obtain the initial representation. This process not only retains the local information of the image but also lays the foundation for subsequent sequencing processing. ViT introduces a special category token, which, together with the image's patch embeddings, constitutes the input of the Transformer model. This category token is essential to the model, and its final state will be used for the final sentiment classification task. Since the Transformer architecture itself does not inherently perceive the order of the sequence, ViT supplements this function by adding positional encoding. The positional encoding provides each patch with spatial position information in the original image, enabling the model to understand the relative positional relationships between different patches. After the image information is flattened and embedded, ViT uses self-attention to process the image features further, enabling the capture of emotional information within the image. Self-attention allows the model to take into account the information from all other patches while processing each patch, thereby capturing the global context information and potential emotional features in the image. This mechanism significant-

Figure 7
Vision Transformer Framework.



ly enhances the model's capacity to understand the global structure of the image.

2.3 Fusion Mechanism

In multi-modal sentiment analysis tasks, processing and fusing features from different data sources is a crucial challenge. To effectively fuse cross-modality features, we employ multi-head attention to address the issue of feature fusion from different data sources.

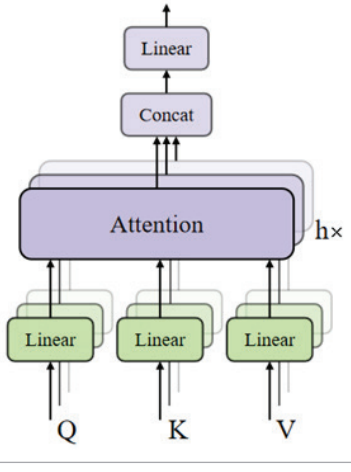
2.3.1 Multi-Head Attention

Self-attention is a technique for computing the correlations between different positions within a sequence in sequence models. The core idea is to allow each element in the sequence to interact with other elements in the sequence to calculate its representation. This mechanism allows the model to capture longer-distance dependencies when processing sequences rather than relying solely on local context from recurrent or convolutional operations. Self-attention generates Q, K, and V matrices through different linear transformations of the input sequence and then calculates the internal attention values of the sequence, which can be further expressed as:

$$Q = XW^{Q},\ K = XW^{K},\ V = XW^{V} \tag{8}$$

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}(\frac{QK^{\text{T}}}{\sqrt{d_k}}) \tag{9}$$

Figure 8
Framework of Multi-Head Attention



Multi-head attention is an extension of traditional self-attention, allowing the model to compute attention weights in parallel across multiple representational subspaces, thereby capturing the relevance of features at different levels of abstraction. Each head of the self-attention focuses on different aspects or combinations of the input features, thereby augmenting the model's cap acity to perceive diverse feature combinations. Under the framework of multi-head attention, the model can perform multiple self-attention operations simultaneously, each using a separate set of parameters. It allows for the parallel extraction of features from different perspectives, and then the outputs of these operations are merged to form an integrated feature representation. This representation can comprehensively capture and fuse features from different modalities, providing rich information for subsequent processing.

2.3.2 Fully Connected Layer

After processing through multi-head attention, the features of images and text are effectively integrated. These integrated features are then refined through a series of FC layers, introducing nonlinear transformations to help the model learn higher-level feature representations. The ultimately integrated features are mapped to sentiment categories through a classifier. Through this comprehensive treatment of multi-head attention and FC layers, the multi-modal sentiment analysis model can more accurately capture and analyze emotional information in cross-modal data, thereby achieving better performance in sentiment recognition tasks.

3. Experiments

3.1 Experimental Environment

1 Introduction to the dataset

We used a total of two datasets. The text dataset used by the first dataset is sourced from the Kaggle platform, which is specifically designed for text sentiment classification tasks and contains approximately 11,000 entries (https://www.kaggle.com/datasets/wcqyfly/bert-data-glue). The records are classified into three sentiment categories: neutral, positive, and negative. In terms of sentiment dis-

tribution, neutral texts account for 47.13% of the total dataset, positive texts account for 37.80%, and negative texts account for 15.07%. To ensure the effectiveness of the model training, establish a ratio of 10:1 for the training set to the test set. In addition, we also used the FER2013 image dataset from the Kaggle platform, which contains over 28,000 training images and more than 3,500 test images (https://www.kaggle.com/datasets/wcqyfly/fer2013-data). The original dataset categorizes images into seven sentiment categories based on facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Each image is a 48×48 pixel grayscale image suitable for facial expression recognition tasks. To match the sentiment categories of the image dataset with the text dataset, we made the necessary adjustments. We unified the expressions of anger, disgust, fear, and sadness into the negative category, marked the expressions of happiness as positive, and kept the label for neutral expressions unchanged. In this way, we ensured consistency in sentiment categories for the image dataset, and we also adjusted the scale of the image data training and test sets to match the text dataset while maintaining the same sentiment distribution. When constructing the negative category image dataset, we paid particular attention to the diversity and consistency of the data distribution. Since the negative category is composed of several original sentiment categories, we randomly extracted according to the original proportion of each negative category image in the FER2013 dataset. Ultimately, in the adjusted negative dataset, the proportion of anger is 29.90%, disgust 3.27%, fear 30.67%, and sadness 36.16%. Through the construction and adjustment of the datasets, we ensured the accuracy and reliability of the experiments, which also contributed to improving the model's capacity for generalization.

The second dataset uses the IMDB dataset as the text corpus, which is an extensively applied movie review dataset for sentiment analysis and NLP tasks (https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews). The IMDB dataset comprises 50,000 user reviews of films, categorized as either positive or negative, with a sentiment category ratio of 1:1. We use 40,000 of these reviews as the training set, while the remaining 10,000 are designated for the test set. This dataset not only provides rich user feedback but also reflects diverse emotional expressions, making it suitable for building and evaluating sentiment analysis models. The image dataset employed is the RAF-DB dataset, which is specifically designed for facial expression recognition and comprises approximately 15,000 RGB images, including around 12,000 images in the training set and about 3,000 in the test set (https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset). This dataset encompasses a diverse range of authentic facial emotional expressions. RAF-DB also categorizes images into seven emotional classes, providing a wealth of emotion labels to support sentiment recognition tasks. In our experiments, we categorized images labeled as happiness into the positive class and grouped anger, disgust, fear, and sadness into the negative class. Additionally, we oversampled the image dataset to match the text dataset. Additionally, during the experiment, both the text dataset and the image dataset underwent specific preprocessing steps. We filtered out some special characters from the text dataset and resized the images in the image dataset to dimensions of 256.

Compared to the first dataset, this dataset is larger in scale and exhibits greater complexity, with longer text lengths and richer information in the image data. Furthermore, there are significant differences in the distributions of the two datasets. Validating the model on both datasets contributes to a more thorough evaluation of the generalization capability of the model proposed in this study.

2 Evaluation metrics

In this experiment, we utilized various evaluation metrics to evaluate the performance of our model thoroughly. These metrics comprise accuracy, precision, recall, F1 score, and AUC value.

Accuracy is the ratio of correctly predicted samples to the total number of samples. In classification tasks, a high accuracy means the model can effectively distinguish between categories.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Precision measures the fraction of samples identified as positive that are truly positive. When the importance of positive samples is high, and there is a

need to reduce false positives, precision becomes a key evaluation metric. A high precision indicates reliability in predicting the positive class.

$$\text{Precision}=\frac{TP}{TP+FP} \tag{11}$$

Recall evaluates the percentage of all actual positive samples the model correctly identifies. In scenarios where it is necessary to identify as many positive samples as possible, recall is essential. A high recall rate indicates that the model can identify a greater number of positive samples, thereby minimizing missed detections.

$$\text{Recall}=\frac{TP}{TP+FN} \tag{12}$$

The F1 score represents the harmonic mean of precision and recall, considering both metrics simultaneously. A high F1 score signifies that the model has effectively maintained a favorable balance between the two.

$$\text{F1 Score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{13}$$

AUC is derived by charting the true positive rate versus the false positive rate at different thresholds, serving as an indicator of the model's performance. A higher AUC value typically reflects superior model performance.

By employing these evaluation metrics comprehensively, we are able to evaluate the model's performance from different perspectives, ensuring that the model not only performs well overall but also meets the expected standards for precision, recall, and overall classification ability. The comprehensive assessment of these metrics helps us to acquire a deeper insight into the model's strengths and weaknesses, offering a more thorough evaluation of its performance.

3 Simulation environment

This experiment used a total of two environments. The first dataset utilized Configuration 1, while the second dataset utilized Configuration 2. The main hardware configuration parameters are shown in Table 1.

Table 1
Experimental Simulation Environment

Parameters	Configuration 1	Configuration 2
OS	Ubuntu 20.04.6 LTS	Ubuntu 20.04.6 LTS
CPU	Intel(R) Xeon(R) CPU	Intel(R) Xeon(R) Gold 6248R CPU
CPU Memory	Memory 29G @ 2.00GHz	Memory 376G @ 3.00GHz
GPU	Tesla P100-PCIE-16GB	NVIDIA A100-PCIE-40GB
Programming Language	Python 3.10.13	Python 3.10.13
Programming environment	PyTorch 2.1.2	PyTorch 2.1.2
CUDA	12.4	12.4

The configuration of a model's hyperparameters is essential to influencing the outcomes of the experimental process. The hyperparameters utilized in this experiment are detailed in Table 2.

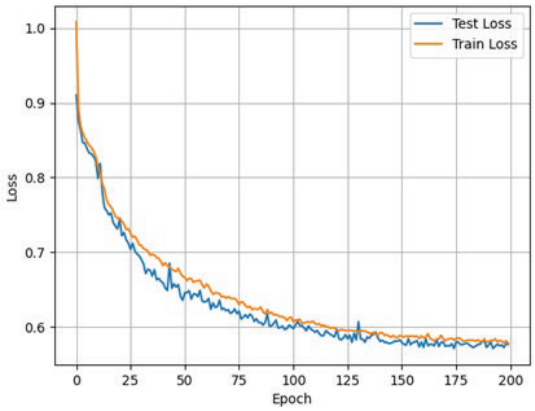
Table 2
Experimental Hyperparameter

Parameters	Dataset 1	Dataset 2
Attention heads	4	4
Optimizer	Adam	Adam
Epochs	200	50
Learning rate	1e-5	1e-5
Weight_decay	0.01	0.01
Image_size	48×48×1	256×256×3
Batch size	16	64

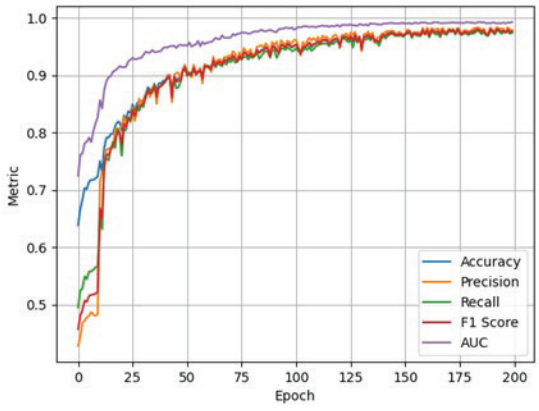
3.2 Model Validation

To assess the effectiveness of the MEA-IFE model, we visualized and analyzed the loss during the training and testing phases, as well as accuracy, precision, recall, F1 score, and AUC metrics during the testing phase, as shown in Figures 9(a)-(b). In Figure 9(a), the orange and blue lines, respectively, record the variations in loss during the training and

Figure 9
Model's LOSS Training Results and Metrics Training Results



(a) Loss



(b) Metrics

testing phases as the number of iterations progresses. The general trend of the loss is relatively stable. At the beginning of the model training, because of the random initialization of parameters, the loss fluctuation is relatively flat in the first 10 iterations. After the 10th iteration, the loss begins to converge rapidly, indicating that the model starts to learn and converge quickly. As the iterations progress, the loss on both the training and testing sets gradually stabilizes after 150 iterations. In Figure 9(b), the blue, orange, green, red, and purple lines, respectively, represent the changes in accuracy, precision, recall, F1 score, and AUC metrics during the testing phase as the number of iterations increases. Among them, the changes in precision, recall, and F1 score can more clearly reflect that the model starts to learn

Table 3
Ablation Study Results.

Model								Metric				
Bert	Bi-LSTM	ResNet	SK-Net	ECA	AlexNet	VIT	Attention	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
√	√	√	×	×	×	×	×	94.01	94.38	94.07	93.77	98.38
√	√	√	√ (Parallel)	×	×	×	×	94.61	95.14	94.67	94.77	98.49
√	√	√	√ (Serial)	×	×	×	×	94.21	94.63	94.27	94.15	98.64
√	√	√	×	√	×	×	×	94.21	94.54	94.22	94.04	98.41
√	√	×	×	×	√	×	×	96.60	96.90	96.36	96.54	99.17
√	√	√	×	×	×	√	×	95.81	96.21	96.19	96.04	98.98
√	√	√	×	×	×	×	√	93.91	93.46	94.17	93.77	97.96
√	√	√	√	√	×	×	×	94.61	94.92	94.85	94.81	98.51
√	√	×	√	√	√	×	×	96.60	96.55	96.68	96.57	99.28
√	√	×	√	√	√	√	×	97.20	97.64	97.16	97.29	99.64
√	√	×	√	√	√	√	√	98.00	98.38	98.00	98.13	99.31

and converge quickly from the 10th round. All metrics remain at a high level after long-term iteration, showing that the model has good generalization capability and stability, thus verifying its effectiveness.

3.3 Ablation Study

To validate the importance of each module in the proposed model, we performed an ablation study using the text dataset from Kaggle and the FER2013 image dataset for the model we proposed. The results of this ablation study are presented in Table 3.

We constructed a baseline model that utilizes a serial architecture of BERT and Bi-LSTM to extract text features. In terms of image feature extraction, ResNet50 is utilized to obtain image features, which are subsequently processed through an FC layer for sentiment analysis. The evaluation metrics we report are all based on the model's performance on the test set. Starting from the Baseline, we incorporated SK-Net, ECA, Multi-head Attention, and Vision Transformer and replaced ResNet50 with AlexNet. The experimental results indicate that, with the exception of Multi-head Attention, the introduction and improvement of the above mechanisms have led to specific improvements in all metrics. For SK-Net, we evaluated both parallel and serial integration modes. The experimental results show that the parallel integration of SK-Net with Bi-LSTM performs better in text feature extraction. Specifically, compared with the serial integration, this parallel integration method achieved improvements of 0.40%, 0.51%, 0.40%, and 0.62% in accuracy, precision, recall, and F1 score, respectively. It is essential to highlight that the introduction of multi-head attention resulted in slightly worse or similar metrics compared to the Baseline. This phenomenon may stem from the limitations in feature extraction of the models used in the Baseline, which could not fully extract features from each modality. Specifically, using only Bert and Bi-LSTM in series in the text feature extraction part could not effectively capture the local features of the text. In the image feature extraction part, ResNet50 tends to focus on local features and neglect global features. Due to the limitations in feature extraction of each modality, multi-head attention could not exert its full advantage when fusing these features.

By introducing SK-Net and Bi-LSTM in parallel integration on the Baseline, the accuracy, precision, recall, F1 score, and AUC metrics improved by 0.60%, 0.76%, 0.60%, 1.00%, and 0.11%, respectively. The introduction of SK-Net enhanced the ability to extract spatial features across multiple scales in the text processing component, compensating for the limitations of Bert and Bi-LSTM feature extraction. With the implementation of the ECA mechanism, the recall rate, F1 score, and AUC improved by 0.18%, 0.04%, and 0.02%, respectively. Changing from ResNet50 to the AlexNet model resulted in improvements of 1.99%, 1.63%, 1.72%, 1.76%, and 0.77% in accuracy, precision, recall, F1 score, and AUC, respectively. The significant performance improvement may be attributed to the relatively high complexity of ResNet50 for the dataset used in the experiment, so using the lighter AlexNet model works better. After adding the Vision Transformer module, the accuracy, precision, recall, F1 score, and AUC increased by 0.60%, 1.09%, 0.48%, 0.72%, and 0.36%, respectively. The results show that introducing VIT on the basis of DCNN allows the model better to balance global and local features during image data processing, thus improving model performance. After removing multi-head attention for feature fusion in the proposed model, accuracy, precision, recall, and F1 score decreased by 0.80%, 0.74%, 0.84%, and 0.84%, respectively. The improved feature extraction module of the model, since the features were effectively extracted, the use of the multi-head attention could effectively improve the model's performance at this point.

To investigate the effect of feature fusion strategies on model performance, we compared the feature fusion mechanism used by MEA-IFE with the late fusion strategy. When the MEA-IFE feature fusion mechanism was replaced with the late fusion method, which only involves weighted fusion of multi-modal outputs from text and images, the model attained an accuracy of only 92.81%, a precision of 93.37%, a recall of 91.27%, an F1 score of 92.23%, and an AUC value of 96.36%, as illustrated in Figure 10. The experimental findings indicate that the feature fusion mechanism employed by MEA-IFE significantly outperforms the late fusion method across all performance metrics, underscoring the importance of feature fusion strategies in enhancing the performance of multi-modal models.

The above ablation study can conclude that the introduction of each module and mechanism has contributed to the model's performance on the test set to varying degrees.

Figure 10
Comparison of Feature Fusion Strategies.

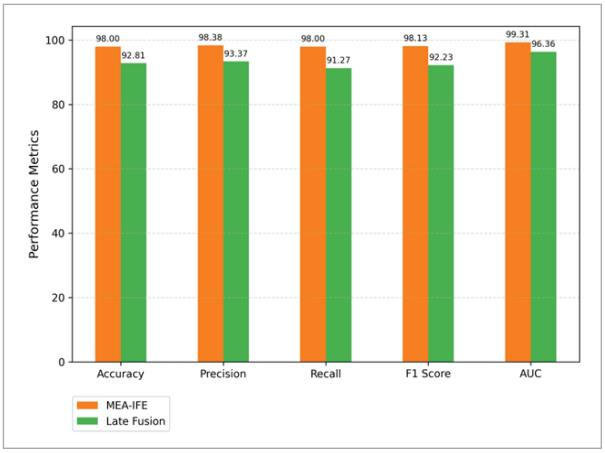
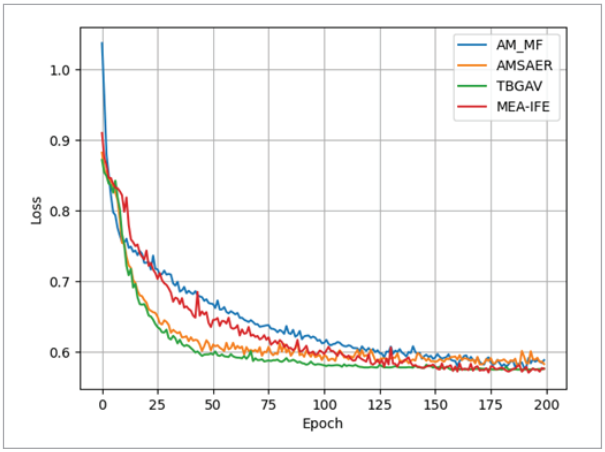


Figure 11
Model Loss Comparison.

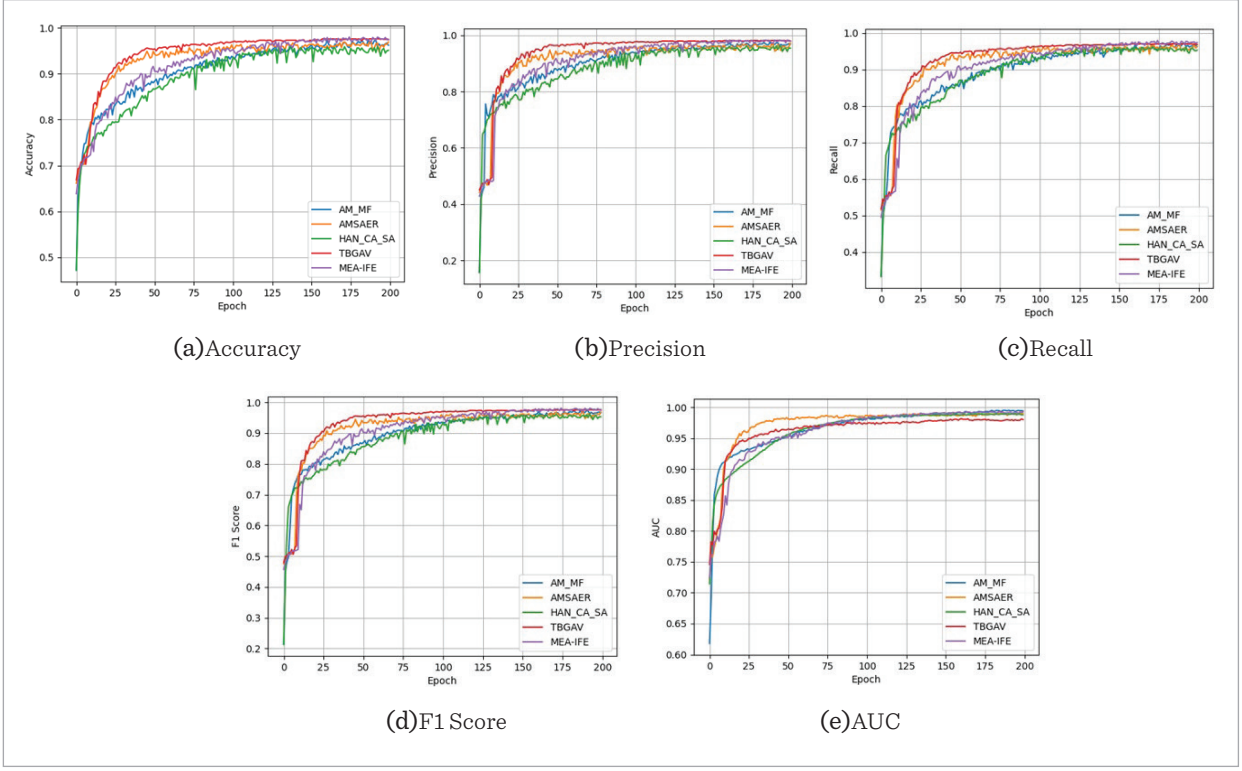


3.4 Comparative experiment

To validate the effectiveness and advantages of the proposed model, we conducted a comparative analysis between the AM-MF [20], AMSAER [3], HAN-CA-SA [16], TBGAV [22], and our proposed MEA-IFE model on the test dataset from Kaggle and the FER2013 image dataset.

Since the HAN-CA-SA model demonstrates a particular difference in loss performance on the dataset rel-

Figure 12
Comparative Experimental Metrics



(a)Accuracy (b)Precision (c)Recall

(d)F1 Score (e)AUC

ative to the other models, we compared the loss differences between MEA-IFE and the AM-MF, AMSAER, and TBGAV models. As shown in Figure 11, it is evident that the MEA-IFE model, due to its higher model complexity and parameter volume, has a relatively slower convergence rate of the loss function during the initial phases of training. However, as the training continues, the MEA-IFE model shows significant improvement, with its loss gradually decreasing and eventually stabilizing at a lower level. It surpasses the AM-MF and AMSAER models and performs slightly better than the TBGAV model.

The comparison of accuracy, precision, recall, F1 score, and AUC metrics between MEA-IFE and AM-MF, AMSAER, HAN-CA-SA, and TBGAV models is shown in Figures 12(a)-(e), with the specific values presented in Table 4.

The comparative results indicate that on the text dataset from Kaggle and the FER2013 image dataset, the proposed MEA-IFE achieved the best performance in terms of accuracy, precision, recall, and F1 score metrics. In terms of the AUC metric, MEA-IFE also ranks high, just behind the AM-MF model. Specifically, for the accuracy metric, MEA-IFE is 0.69%, 0.89%, 1.89%, and 0.40% better than the AM-MF, AMSAER, HAN-CA-SA, and TBGAV models, respectively. For the precision metric, the model leads the AM-MF, AMSAER, HAN-CA-SA, and TBGAV models by 0.68%, 1.13%, 2.11%, and 0.20%, respectively. In terms of recall, the model is 0.75%, 0.88%, 1.79%, and 0.92% better than the AM-MF, AMSAER, HAN-CA-SA, and TBGAV models, respectively. Additionally, for the F1 score metric, the model also leads the AM-MF, AMSAER, HAN-CA-SA, and TBGAV models by 0.82%, 1.11%, 1.93%, and 0.52%, respectively.

To further validate the model's robustness against noise and outliers, we processed the data by introducing noise and anomalies to the original dataset. Specifically, we randomly replaced 1% of the entries in the text dataset with empty strings to evaluate the model's performance. The detailed experimental results are presented in Table 5. As illustrated in the table, despite the presence of noise, MEA-IFE maintains a high level of performance comparable to that observed with noise-free data and generally outperforms other models across various metrics. It provides strong evidence of MEA-IFE's robustness in the presence of noise.

To evaluate the model's ability to generalize across varying data distributions, we conducted further ex-

Table 4
Comparative Experimental Results without Noise on Dataset 1.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
HAN-CA-SA [16]	96.11	96.27	96.21	96.20	98.93
AMSAER [3]	97.11	97.25	97.12	97.02	99.13
AM-MF [20]	97.31	97.70	97.25	97.31	**99.53**
TBGAV [22]	97.60	98.18	97.08	97.61	98.17
MEA-IFE	**98.00**	**98.38**	**98.00**	**98.13**	99.31

Table 5
Comparative Experimental Results with Noise on Dataset 1.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
HAN-CA-SA [16]	95.60	96.27	95.67	95.86	98.84
AMSAER [3]	97.30	97.30	96.98	96.87	98.98
AM-MF [20]	94.31	94.97	94.96	94.86	99.09
TBGAV [22]	97.60	**98.24**	97.17	97.69	98.13
MEA-IFE	**98.00**	98.00	**97.92**	**97.86**	**99.36**

Table 6
Comparative Experimental Results on Dataset 2.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
HAN-CA-SA [16]	95.04	95.04	95.04	95.04	98.77
AMSAER [3]	96.13	96.14	96.13	96.13	99.01
AM-MF [20]	96.02	96.02	96.02	96.02	99.05
TBGAV [22]	93.58	93.58	93.56	93.58	97.58
MEA-IFE	**96.33**	**96.34**	**96.33**	**96.33**	**99.10**

periments using the IMDB text dataset and the RAF-DB image dataset. This dataset exhibits significant distribution differences compared to the first dataset and is larger in scale and greater in complexity. The experiments on this dataset effectively validate the model's generalization ability and its capacity to address complex sentiment analysis challenges. The comprehensive results are provided in Table 6.

In the IMDB text dataset and the RAF-DB image dataset, the MEA-IFE model achieved an accuracy of 96.33%, surpassing that of the HAN-CA-SA (95.04%), AMSAER (96.13%), TBGAV (93.58%), and AM-MF (96.02%) models. It indicates that MEA-IFE effectively captures underlying patterns within the data, demonstrating its robustness in complex sentiment analysis tasks. Additionally, its precision, recall, F1 score, and AUC are all superior to those of the other models. The experimental results indicate that MEA-IFE demonstrates significant practical application potential and good generalization capability in addressing complex sentiment analysis challenges.

In summary, the comparative experiments have proven the efficacy and advantages of the proposed MEA-IFE model in the field of sentiment analysis.

4. Discussion

4.1 Summary of Work

The MEA-IFE model presented in this study significantly improves the accuracy of sentiment analysis by integrating text and image information and optimizing feature extraction and fusion mechanisms. The model is characterized by high predictive accuracy, strong feature extraction capabilities, and adaptive fusion, making full use of features from different modalities. For text feature extraction, MEA-IFE introduces an SK-Net in parallel with BERT-BiLSTM, enhancing the extraction of multi-scale spatial features and employing the ECA mechanism to strengthen the recognition of important textual information. For image feature capture, the model introduces Vision Transformer on the basis of using DCNN for feature extraction to improve the model's global representation capability.

4.2 Differences from Other Works

The primary distinction between the MEA-IFE model and existing work resides in its innovative multi-modal fusion strategy and feature extraction methods. Compared with single-modal sentiment analysis models [14], MEA-IFE can more comprehensively capture emotional information, avoiding biases from single-source information. Furthermore, compared with existing multi-modal models [10], MEA-IFE enhances the capture of local and global features through the introduction of SK-Net and Vision Transformer and strengthens the recognition of feature importance through the ECA mechanism.

4.3 Limitations and Improvements

Despite the notable advancements made by the MEA-IFE model in sentiment analysis, it does exhibit several limitations. With a parameter count of 215.6M, it surpasses that of AM-MF (213.0M), TBGAV (135.1M), HAN-CA-SA (113.0M), and AMSAER (170.0M). This higher complexity and resource demand may lead to increased training costs and challenges during the training process. Additionally, while the MEA-IFE model excels in multi-modal feature fusion, it may struggle with certain nuanced emotional expressions. The model's generalization capabilities across diverse datasets also require further validation and en-

hancement. Nonetheless, the MEA-IFE model holds significant potential for various applications, such as sentiment analysis in social media, product reviews, public safety monitoring, and enhancing customer service experiences. Further optimization shows promise for multi-language sentiment analysis and cross-cultural emotion recognition.

4.4 Error Case Analysis

Despite the significant progress made by MEA-IFE in sentiment analysis, misjudgments still tend to occur when dealing with texts that express complex emotions. For example, consider the following text: "Ok, this movie is stupid. I mean that in a good way, however. It was stupid on purpose and was one of the better 'stupid' movies I have seen. The jokes and gags are purposefully bad but delivered in a way that struck all the right notes with me. The supporting characters were pretty shallow and mediocre. There is a pretty weak plot, but it works just fine." In this case, MEA-IFE might incorrectly assess the sentiment as negative.

There are several potential reasons for this misjudgment. First, the text contains words such as "stupid," "shallow," and "weak," which are typically categorized as negative vocabulary. Although the overall sentiment is positive, references to the flaws in supporting characters and the plot might lead the model to lean towards a negative sentiment assessment. Second, the model may fail to grasp that the term "stupid" is used in an appreciative context, indicating a lack of sensitivity to humor and irony. Additionally, the training data may lack instances of specific emotional expressions like "deliberately stupid", limiting the model's ability to recognize such nuanced sentiments. Therefore, MEA-IFE still exhibits certain limitations when addressing complex emotional expressions, particularly in lengthy and intricate texts.

4.5 Future Prospects

In response to the limitations of the MEA-IFE model, future work can explore the following aspects: Firstly, explore more lightweight network structures or model pruning techniques to reduce the model's parameter volume and computational costs. Secondly, research on improved models of BERT, such as Roberta or ALBERT, to increase the efficiency of text feature extraction. In addition, more advanced image processing techniques and network structures can be introduced to enhance the model's capacity to handle complex image scenarios. Finally, research on more efficient feature fusion techniques is needed to balance the model's performance and computational resource requirements.

5. Conclusion

This paper presents an innovative multi-modal sentiment analysis framework, MEA-IFE, characterized by its effective feature extraction capabilities and high predictive accuracy. The model addresses potential information loss and expression limitations in the BERT-BiLSTM text feature extraction process by introducing a parallel structure of SK-Net and BiLSTM, enhancing the model's capacity to capture multi-dimensional text features. Additionally, to improve the precision of text feature extraction, the model integrates the ECA mechanism, which helps the model to capture key information in the text more keenly. In terms of image processing, to address the issue of DCNN potentially overlooking global features, the MEA-IFE model introduces Vision Transformer, combining CNN and Transformer to enhance the capture of both global and detailed image features. At the critical stage of feature fusion, the MEA-IFE model employs multi-head attention, achieving dynamic fusion of text and image features, profoundly exploring the interactive potential across various data modalities, thus significantly enhancing the performance of sentiment analysis tasks.

The proposed model was validated on a text dataset from Kaggle and the FER2013 image dataset. Experimental results indicate that the proposed model achieves high performance, with an accuracy rate of 98.00%, precision of 98.38%, recall rate of 98.00%, F1 score of 98.13%, and an AUC indicator of 99.31%. Ablation experiments indicate that the introduction of each module effectively improves the metrics, thereby verifying the effectiveness of the improvement strategies. When compared with models such as AM-MF, AMSAER, HAN-CA-SA, and TBGAV, the results show that the proposed model attained the highest performance regarding accuracy, precision, recall, and F1 score, with respective improvements of 0.40%, 0.20%, 0.75%, and 0.52%, and it also ranks high in the AUC indicator. The MEA-IFE model pre-

sented in this study exhibits high predictive accuracy and powerful feature integration capabilities, meeting the demand for high precision in complex multi-modal sentiment analysis tasks. Considering the complexity and large number of parameters in the proposed model's structure, future work could explore lighter networks and further optimize the network structure to reduce the model's parameter volume while maintaining its performance. In terms of feature fusion, different fusion strategies can be explored to enhance the interaction and fusion of features from different modalities.

References

1. Abdullah, T., Ahmet, A. Deep Learning in Sentiment Analysis: Recent Architectures. ACM Computing Surveys, 2022, 55(8), 1-37. https://doi.org/10.1145/3548772
2. Ameer, I., Bölücü, N., Siddiqui, M. H. F., Can, B., Sidorov, G., Gelbukh, A. Multi-Label Emotion Classification in Texts Using Transfer Learning. Expert Systems with Applications, 2023, 213, 118534. https://doi.org/10.1016/j.eswa.2022.118534
3. Aslam, A., Sargano, A. B., Habib, Z. Attention-Based Multi-Modal Sentiment Analysis and Emotion Recognition Using Deep Neural Networks. Applied Soft Computing, 2023, 144, 110494. https://doi.org/10.1016/j.asoc.2023.110494
4. Bashiri, H., Naderi, H. Comprehensive Review and Comparative Analysis of Transformer Models in Sentiment Analysis. Knowledge and Information Systems, 2024, 1-57. https://doi.org/10.1007/s10115-024-02214-3
5. Basiri, M. E., Nemati, S., Abdar, M., Nemati, S., Abdar, M., Cambria, E., Acharya, U. R. ABCDM: An Attention-Based Bidirectional CNN-RNN Deep Model for Sentiment Analysis. Future Generation Computer Systems, 2021, 115, 279-294. https://doi.org/10.1016/j.future.2020.08.005
6. Brauwers, G., Frasincar, F. A Survey on Aspect-Based Sentiment Classification. ACM Computing Surveys, 2022, 55(4), 1-37. https://doi.org/10.1145/3503044
7. Campos, V., Jou, B., Giro-i-Nieto, X. From Pixels to Sentiment: Fine-Tuning CNNs for Visual Sentiment Prediction. Image and Vision Computing, 2017, 65, 15-22. https://doi.org/10.1016/j.imavis.2017.01.011
8. Cao, J., Chen, J., Han, J., Li, H. Sentiment Classification of Image Based on Adaboost-BP Neural Network. Journal of Shanxi University, 2013, 36(3), 331-337.
9. Chan, J. Y. L., Bea, K. T., Leow, S. M. H., Phoong, S. W., Cheng, W. K. State of the Art: A Review of Sentiment Analysis Based on Sequential Transfer Learning. Artificial Intelligence Review, 2023, 56(1), 749-780. https://doi.org/10.1007/s10462-022-10183-8
10. Das, R., Singh, T. D. Multi-Modal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. ACM Computing Surveys, 2023, 55(13s), 1-38. https://doi.org/10.1145/3586075
11. D'Mello, S. K. A Review and Meta-Analysis of Multi-Modal Affect Detection Systems. ACM Computing Surveys, 2015, 47(3), 1-36. https://doi.org/10.1145/2682899
12. Jiang, X., Song, C., Xu, Y., Li, Y., Peng, Y. Research on Sentiment Classification for Netizens Based on the BERT-BiLSTM-TextCNN Model. Peer Journal of Computer Science, 2022, 8, e1005. https://doi.org/10.7717/peerj-cs.1005
13. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2012, 25. https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
14. Liu, X., Shi, T., Zhou, G., Liu, M., Yin, Z., Yin, L., Zheng, W. Emotion Classification for Short Texts: An Improved Multi-Label Method. Humanities and Social Sciences Communications, 2023, 10(1), 1-9. https://doi.org/10.1057/s41599-023-01816-6
15. Rezaeinia, S. M., Rahmani, R., Ghodsi, A., Veisi, H. Sentiment Analysis Based on Improved Pre-Trained Word Embeddings. Expert Systems with Applications, 2019, 117, 139-147. https://doi.org/10.1016/j.eswa.2018.08.044 Sujeesha, A. S., Mala, J. B., Rajan, R. Automatic Music Mood Classification Using Multi-Modal Attention Framework. Engineering Applications of Artificial Intelligence, 2024, 128, 107355. https://doi.org/10.1016/j.engappai.2023.107355
16. Wang, D., Guo, X., Tian, Y., Liu, J., He, L., Luo, X. TETFN: A Text Enhanced Transformer Fusion Network for Multi-Modal Sentiment Analysis. Pattern Recognition, 2023, 136, 109259. https://doi.org/10.1016/j.patcog.2022.109259

17. Wang, G., Shin, S. Y., Lee, W. J. A Text Sentiment Classification Method Based on LSTM-CNN. Journal of The Korea Society of Computer and Information, 2019, 24(12), 1-7. https://doi.org/10.9708/jksci.2019.24.12.001

18. Wang, K., Zheng, Y., Fang, S., et al. Long Text Aspect-Level Sentiment Analysis Based on Text Filtering and Improved BERT. Journal of Computer Applications, 2020, 40(10), 2838-2844. https://doi.org/10.11772/j.issn.1001-9081.2020020164

19. Xie, S., Li, J. A Multi-Modal Sentiment Analysis Method Integrating Multi-Layer Attention Interaction and Multi-Feature Enhancement. International Journal of Intelligent Transportation Systems and Applications, 2024, 17(1), 1-20. https://doi.org/10.4018/IJITSA.335940

20. Yadav, A., Vishwakarma, D. K. Sentiment Analysis Using Deep Learning Architectures: A Review. Artificial Intelligence Review, 2020, 53(6), 4335-4385. https://doi.org/10.1007/s10462-019-09794-5

21. Zhang, K., Wang, S., Yu, Y. A TBGAV-Based Image-Text Multi-Modal Sentiment Analysis Method for Tourism Reviews. International Journal of Information Technology and Web Engineering, 2023, 18(1), 1-17. https://doi.org/10.4018/IJITWE.334595

22. Zhang, L., Wang, S., Liu, B. Deep Learning for Sentiment Analysis: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4), e1253. https://doi.org/10.1002/widm.1253

23. Zhao, T., Meng, L., Song, D. Multi-Modal Aspect-Based Sentiment Analysis: A Survey of Tasks, Methods, Challenges and Future Directions. Information Fusion, 2024, 112, 102552. https://doi.org/10.1016/j.inffus.2024.102552

24. Zhu, L., Zhu, Z., Zhang, C., Xu, Y., Kong, X. Multi-Modal Sentiment Analysis Based on Fusion Methods: A Survey. Information Fusion, 2023, 95, 306-325. https://doi.org/10.1016/j.inffus.2023.02.028