Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9940817, 10 pages
https://doi.org/10.1155/2022/9940817

Research Article

Sentiment Classification of Tourism Reviews Based on Visual and Textual Multifeature Fusion

Shujun Wei[1] **and Song Song**[2]

[1]*Business School, University of Chinese Academy of Social Science, Beijing 102445, China*
[2]*School of Computer Science, Wuhan University, Wuhan 430072, China*

Correspondence should be addressed to Shujun Wei; b19500167@ucass.edu.cn

Received 11 March 2022; Revised 1 April 2022; Accepted 25 April 2022; Published 29 May 2022

Academic Editor: Deepak Kumar Jain

Most of the traditional tourism review sentiment analysis methods ignore the complementary role between review text and images and do not pay attention to the differences in the sentimental expression of different users and the characteristics of the text content of tourism reviews other than the plain text such as emojis. For this reason, a sentiment analysis method for tourism online reviews based on multifeature fusion of images and text is proposed. Firstly, a text sentiment classification model is constructed, and a variety of sentiment features are combined to form a multi-input matrix, and then, it is input into a multichannel CNN (convolutional neural network) to extract sentiment features in order to complete the text sentiment classification. In addition, an image sentiment classification model is constructed by merging the global image and the face image. On the basis of the CNN, the supervision module with weighted loss is added to extract the facial sentiment features, and the facial target sentiment is fused with the sentiment directly recognized by the whole image, and the sentiment polarity of the image in the posted tourism review is determined. Finally, a decision fusion method is designed to fuse the output of the text and image sentiment classification models. The experimental results show that the proposed image-text fusion sentiment classification model effectively enhances the ability of the model to capture sentimental semantics of tourism reviews through the combination of text content features and image sentiment features and achieves excellent results in multiple performance metrics, with better sentiment classification performance than other state-of-the-art models.

1. Introduction

The concept of smart tourism has recently received a lot of attention from academics and practitioners. The concept was aimed at accelerating service innovation, improving the tourism experience, and enhancing the competitiveness of tourism destinations through the development of Internet, communication, big data, and other technology as a basis [1]. Sentiment analysis, the frontier of social media analytics, is widely used to target users for product marketing, political forecasting, stock prediction, and mental health analysis [2]. In recent years, with the widespread use of mobile photography devices and the continuous progress of the network environment, graphic-text user reviews have become quite common on various social media platforms and travel e-commerce website platforms, and a large number of users share their travel experiences online, and travel reviews and photos account for a large proportion of online information and are constantly being added or updated, providing new research opportunities and challenges for multimedia and data mining research and applications [3].

More than half of travelers plan their own itineraries by browsing reviews when choosing attractions or hotel accommodations, and online reviews are more likely to influence travel planning than recommendations from friends or family members. Users are free to express their thoughts and opinions on the Internet, so travelling websites generate a large number of reviews about tourist attractions, and there is a lot of potential value in the reviews [4]. The government and tourism industry can mine the reviews and analyze their results to facilitate relevant departments and scenic spots to identify problems, improve visitor services, and increase visitor satisfaction. Therefore, accurate mining of tourists' sentiments behind the reviews has positive guiding significance for the development of the industry [5].

Tourist sentiment comments published by tourists on the Internet platform are characterized by randomness and abstraction [6]. The existing feature fusion methods have little research on the natural connection between the two modalities of images and text. As described by Chen et al. [7], images and text can not only generate entity-level correspondences but also reflect sentimental consistency through the visual features at the bottom and middle layers.

This paper proposes a multifeature fusion method for sentiment analysis of graphic-text tourism reviews. The main contributions of this paper are listed as follows: (1) a multichannel CNN (convolutional neural network) is used to learn and extract more comprehensive sentiment semantics from multi-input embeddings, the combination of different features not only forms new features but also allows the features to be related and influenced with each other. In addition, the text features are matched with the tag features to effectively utilize the sentiment information contained in the text and tags and the connections between them; (2) an image sentiment classification model based on the fusion of global image and face image is designed; and (3) finally, the outputs of text and image sentiment classification models are merged by the fusion module to achieve accurate sentiment classification of online tourism reviews.

The remainder of this paper is organized as follows. Section 2 introduces the research related to sentiment analysis of online travel reviews. The proposed multifeature fusion graphical and textual travel review sentiment classification model is explained in Section 3. Section 4 gives the experimental results and discussions. Finally, Section 5 concludes the whole paper.

2. Related Research

Due to the successful application of deep learning models in the field of NLP (natural language processing), more and more research has made great progress in applying deep learning techniques to sentiment classification tasks. Kalchbrenner et al. [8] proposed a method to judge the sentiment polarity of tweets using CNN. Wang et al. [9] adopted LSTM (long short-term memory neural network) to analyze the sentiment categories of text. The above studies initially explored the application of deep learning in sentiment analysis and achieved better classification performance than traditional classifiers without the need for manual feature extraction. Online travel reviews contain unique data that traditional text does not have, such as emojis and tags. Inspired by this, recently more and more researchers have been studying how to make full use of some of the data in online travel reviews to improve sentiment classification performance. Vo and Zhang [10] proposed to use diverse feature information to improve the accuracy of tweet sentiment classification.

Most of the existing sentiment analysis research on social platforms is mainly based on text. With the rise of image sharing applications, some sentiment analysis methods for images have also been proposed. Campos et al. [11] used convolutional networks to extract features from images, and passed the extracted high-level semantic features through feedforward layers for sentiment classification, and achieved higher accuracy than traditional machine learning (ML) methods. But the sentiment of an image is not only related to high-level semantics. Li et al. [12] proposed a multilayer deep representation network, which connected the different convolutional layers of AlexNet to extract the image features of each layer, and then, these features were concatenated to effectively utilize different levels of image sentiment information.

However, due to the diversity of online tourism review data, only images or text cannot fully reveal the sentiment of different users, so researchers have also begun to study sentiment analysis of online reviews based on information fusion. You et al. [13] claimed that images and text should be analyzed jointly in a structured manner, and a semantic tree structure was established based on syntactic analysis to map word and image regions in text to achieve sentiment classification from visual-textual fusion.

Majumder et al. [14] proposed a context-aware high-dimensional fusion modeling method, in which modalities such as text, sound, and pictures are fused multiple times to obtain fused features. However, based on the principle of the attention mechanism, the emotion that can reflect the text, picture, or audio is only concentrated on some words in the text or some regions in the picture and audio, and the fusion of the overall information of different modalities will introduce noise and cause information redundancy. Huang et al. [15] proposed a deep multimodal fusion method. After modeling text and images based on the attention mechanism, the generated text and image vectors with key features are fused and input into the fully connected layers to perform sentiment predictions. However, in reality, there is a corresponding relationship between a certain keyword that reflects the sentiment of the text and certain parts of the image, and this method does not take this relationship into consideration.

3. Tourism Review Sentiment Analysis Model Based on Multifeature Fusion of Images and Text

The framework of the multifeature fusion sentiment analysis method for graphic-text tourism reviews proposed in this paper is shown in Figure 1. According to the different functions, the framework is divided into three modules: text sentiment classification module based on multifeature fusion, image sentiment classification module based on the combination of global image and face image, and image-text sentiment fusion module.

3.1. Text Sentiment Classification Based on Multifeature Fusion.

In the sentiment classification task of tourism review text, the extraction of sentiment features is particularly important for the classification results. This paper proposes a multifeature fusion text sentiment classification model, which combines different features as the multi-input of the

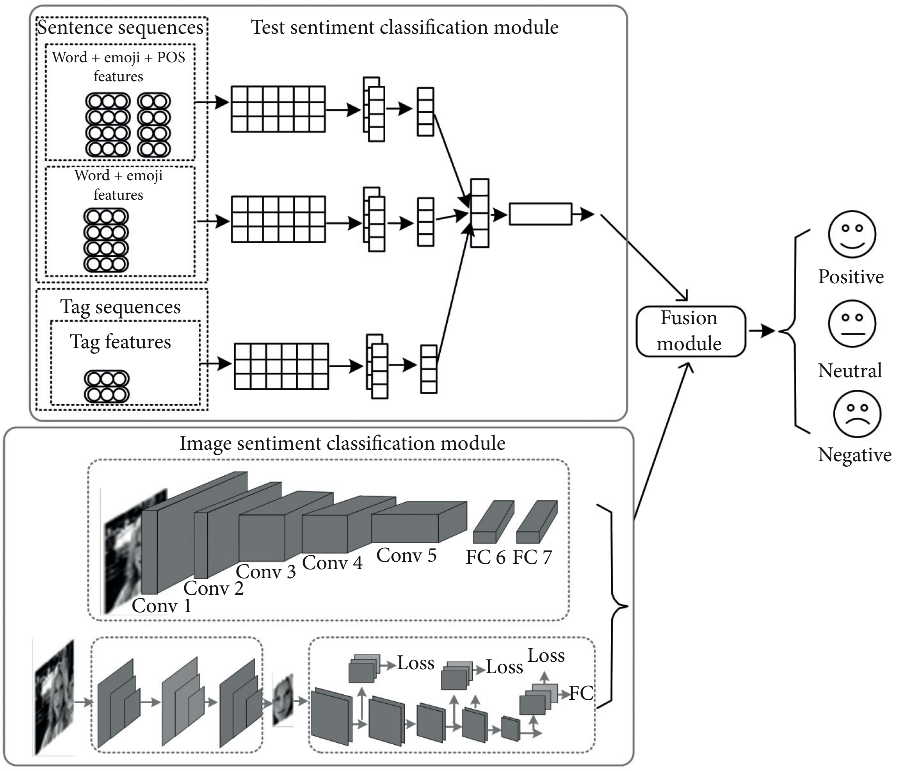


Figure 1: Image-text sentiment classification framework.

CNN. This enables the model to obtain more comprehensive sentiment information and achieve better sentiment classification results. The tourism text classification model is shown in Figure 2.

3.1.1. Embedding Construction. Four kinds of features are extracted from tourism review texts: word features, POS (part-of-speech) features, emoji features, and tag features. In word features, taking the word of the sentence as the unit, each word is mapped to a word vector. There are two columns in the word vector dictionary, one column contains the words, and the other includes the distributed word vector representation corresponding to the words.

Let N be the number of elements in the dictionary, d is the dimension of each word vector, then $\boldsymbol{M} \in R^{d\times N}$ is the representation matrix of the entire word vector dictionary.

For a sentence sequence $S = \{w_1, w_2, \cdots, w_n\}$, by looking up the word vector corresponding to each word w_i in the word vector matrix $R^{d\times N}$, and concatenate the word vector corresponding to each word in each sentence sequence in turn, the word vector matrix of the entire sentence sequence is obtained:

$$\boldsymbol{M}_S = \vec{w}_1 \oplus \vec{w}_2 \oplus \cdots \oplus \vec{w}_n, \tag{1}$$

where $\vec{w}_i$ denotes word vector, and $\oplus$ represents vector concatenation operator.

In POS features, Hownet sentiment word list [16] is used to update POS labels of special words in sentences. In the process of feature extraction, the feature information of words that play an important role in sentiment classification can be preserved. The POS labels are vectorized and mapped to a distributed POS vector:

$$\boldsymbol{Pos}_{1:n} = \boldsymbol{pos}_1 \oplus \boldsymbol{pos}_2 \oplus \cdots \oplus \boldsymbol{pos}_n, \tag{2}$$

where $\boldsymbol{pos}_i$ denotes the POS vector, and i represents the ordinal number of POS vector.

In emoji features, sentences with sentimental emojis are easier to determine sentiment tendencies, and sentimental emojis have a stronger sentiment indicating effect than sentiment words. For each sentence, if there is an emoji, convert the emoji into a distributed vector and find its corresponding sentiment polarity in Hownet:

$$\boldsymbol{M}_E = \vec{\boldsymbol{e}}_1 \oplus \vec{\boldsymbol{e}}_1 \oplus \cdots \oplus \vec{\boldsymbol{e}}_E, \tag{3}$$

where $\vec{\boldsymbol{e}}_i$ denotes the vector of the emoji e_i, E is the number of emojis in the sentence.

In tag features, in order to make full use of the emotional information expressed by #tags, positive tags, negative tags, and no-sentiment tags are converted into distributed vectors corresponding to positive, negative, and neutral sentiments, respectively. The vectorized tags are denoted by $\boldsymbol{Tag}$, where m is the number of categories of tags, and $\boldsymbol{tag}_i$ is the vector of tag y_i.

3.1.2. Text Sentiment Classification CNN. The four types of features are combined and transformed to construct three new combined features: (1) word features+emoji features+

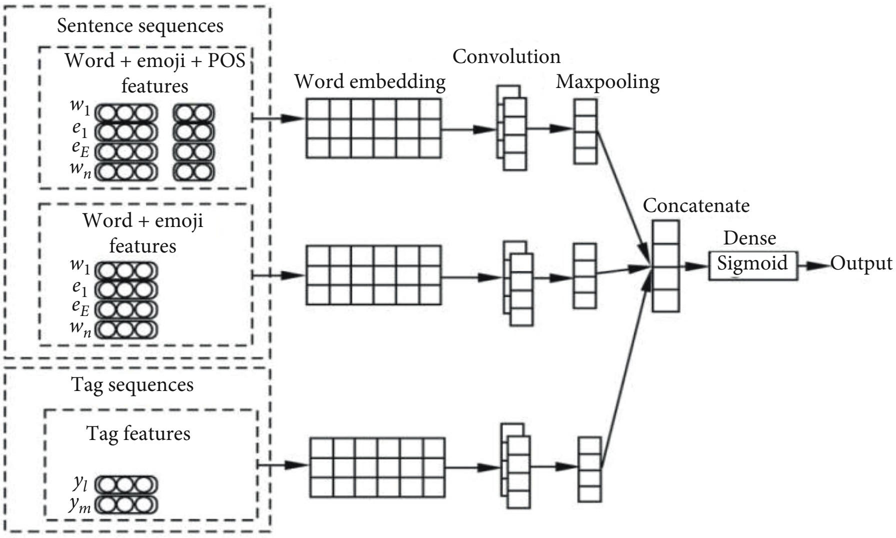


Figure 2: Tourism review text sentiment classification model.

POS features; (2) word features+emoji features; and (3) tag features. Then, the combined features are put through the three channels of CNN, respectively. The word vectors, emoji vectors, and POS vectors are concatenated to form the first embedding matrix:

$$\boldsymbol{V}_1 = \left\{ \vec{\boldsymbol{w}}_1 \oplus \boldsymbol{pos}_1, \vec{\boldsymbol{e}}_1 \oplus \boldsymbol{pos}_2, \cdots, \vec{\boldsymbol{w}}_n \oplus \boldsymbol{pos}_{n+E} \right\}. \tag{4}$$

The word vectors and the emoji vectors are concatenated to form the second embedding matrix:

$$\boldsymbol{V}_2 = \left\{ \vec{\boldsymbol{w}}_1, \cdots, \vec{\boldsymbol{e}}_1, \cdots, \vec{\boldsymbol{e}}_E, \cdots, \vec{\boldsymbol{w}}_n \right\}. \tag{5}$$

The third embedding matrix is represented by tag vectors:

$$\boldsymbol{V}_3 = \{ \boldsymbol{tag}_1, \boldsymbol{tag}_2, \cdots, \boldsymbol{tag}_m \}. \tag{6}$$

After that, convolution operations are performed on the input features using convolution kernels with different sizes to extract rich abstract features in the sentence:

$$c_i = f(\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h-1} + b), \tag{7}$$

where $w \in R^{h\times m}$ represents the weight of the convolution kernel and $h \times m$ represents the size of the convolution kernel window; b is the bias, f denotes activation function, and Relu is used as the activation function for the convolution layer; and $\boldsymbol{x}_{i:i+h-1}$ represents the word embedding matrix from the word to the $(i+h-1)$th word in the sentence sequence. c_i represents the ith eigenvalue obtained by the convolution operation.

After the convolution operation, each sentence can get a feature map c:

$$\boldsymbol{c} = [c_1, c_2, \cdots, c_{n-h+1}], \tag{8}$$

where n represents the length of the sentence sequence.

The pooling layer performs sampling operations by setting a fixed step size in the pooling area, and maximum pooling is used to perform downsampling operations on the convolved feature maps c in different channels to extract the most important features in each channel. The final feature matrix $\hat{\boldsymbol{c}}$ extracted from the text sequence by m convolution kernels is expressed as

$$\hat{\boldsymbol{c}} = [\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_m], \tag{9}$$

where m is the number of convolution kernels.

After convolution and pooling, the features of the three channels are represented by $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{Y}_j$, respectively, which are combined to form the final vector and input to the fully connected layer $M_{3m\times 1}$ to output the sentiment classification result:

$$s_j = \sigma\left(M_{3m\times 1}\left(\boldsymbol{X}_1 \oplus \boldsymbol{X}_2 \oplus \boldsymbol{Y}_j\right)\right), \tag{10}$$

where σ represents the sigmoid function. After the nonlinear function transformation, a matching score between 0 and 1 is generated between the text and each sentiment. The larger the score value, the closer the sentiment category of the text is to the sentiment (positive, neutral, or negative).

3.2. Tourism Review Image Sentiment Classification Module. In the proposed framework, the facial target and the whole image sentiment recognition models are integrated to determine the sentiment polarity of images in tourism reviews. The details are described below.

3.2.1. Sentiment Analysis of the Global Image. CNN have performed well in computer vision-related tasks, with the advantage of enabling feature extraction and classification in an end-to-end manner. The whole image sentiment analysis uses the VGG16 network structure to extract the features of the image, of which 16 means that the structure consists of 13 convolutional layers and 3 fully connected layers. Since training a CNN model requires a large number of datasets, and the number of existing labeled image sentiment datasets is relatively small, the model parameters pretrained on the ImageNet dataset are used to transfer the trained convolutional layer (conv1~conv5) parameters to in VGG16. In order to speed up the training, the last two fully connected layers FC6 and FC7 are replaced by convolutional layers composed of convolution kernels 7 × 7 and 1 × 1, respectively. Since this paper studies the three-category problem of sentiment classification for tourism reviews, the output dimension of FC7 is modified to 3, and finally, the sentiment recognition model of the entire image is obtained by fine-tuning, as shown in Figure 3. The learning rate is 0.001, the momentum is set to 0.9, the activation function is Relu, and SGD (stochastic gradient descent) is used to update the network parameters. In order to prevent over fitting, a dropout layer is added after FC6, and its coefficient is set to 0.5.

Let the training sample dataset be $\{(I_i,\beta_i\}\}_{i=1}^{N}$, where N is the size of the training set, I_i is the input image, and β_i is the corresponding sentiment label. The cross-entropy loss function ∂_{global} of the global image sentiment recognition model is expressed as follows:

$$\partial_{global}(W) = \sum_{i=1}^{N}\sum_{j\epsilon\beta} q(\beta_i, j)\log p\left(\beta_i = j \middle| d_j, w_j\right), \tag{11}$$

where W is the parameter set of model, and W_j is the linear mapping parameter. If $\beta_i = j$, then indicator function $q(\beta_i, j) = 1$; otherwise, $q(\beta_i, j) = 0$. d_i is the output of FC7. After linear mapping, we can get the sentiment polarity probability of the global image with the softmax function.

3.2.2. Face Target Sentiment Analysis. In order to obtain the sentiment polarity of the face target, firstly, the MTCNN [17] algorithm is used to obtain the accurate face target position, and then, a face target sentiment recognition model is constructed to extract the face target sentiment features. Finally, the sentiment polarity of the face target is obtained. The face target sentiment analysis framework is shown in Figure 4.

(1) Face Target Detection. MTCNN is a deep learning cascading framework for face detection, which consists of three neural networks with different degrees of complexity, namely, P-Net, R-Net, and O-Net. In the training phase, in order to speed up the training convergence, each neural network inputs face images of different sizes. The input of P-Net is randomly cut to generate 12 × 12 size images. Then, the candidate face images are obtained through the P-Net network, and the image sizes are adjusted to 24 × 24 and input to the R-Net network for refinement. Finally, the face images obtained by the R-Net network are adjusted to a 48 × 48 and input into the O-Net to obtain the face bounding boxes and key point information. Different from the training phase, in the testing phase, in order to detect faces of different sizes, the image pyramid is used to scale the image to different sizes and then input to P-Net, R-Net, and O-Net in turn to get the location information of the face target.

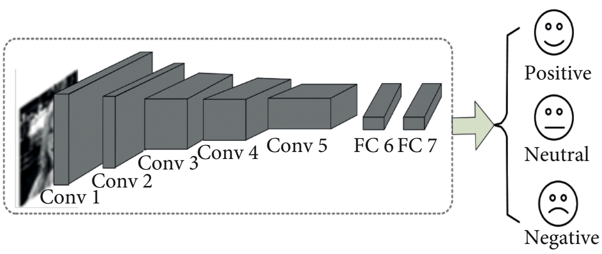


Figure 3: Global image sentiment classification network.

(2) Face Target Sentiment Recognition. Based on SSE (supervised scoring ensemble) algorithm [18], the supervision module is introduced into VGG16, and different weights are applied to the losses of different supervision modules to construct a facial target sentiment recognition model. Since the feature map of the output layer is learned from the feature map of the hidden layers (shallow layers and intermediate layers), the shallow supervision module SS_block, the intermediate supervision module IS_block, and the deep supervision module DS_block are added at the end of conv1_2, conv3_3, and conv5_3, respectively, so that it can not only supervise the output layer of CNN but also supervise the facial target sentiment feature information in the middle and shallow layers at the same time, so as to improve the quality of sentiment feature mapping of the output layer and ensure the accuracy of the final salient target sentiment recognition. In order to avoid the loss of useful emotional features in the pooling layer, 7 × 7 convolutional layers with stride 2 are used.

The SS_block and IS_block modules can supervise the learning of face features from the shallow and middle layers. In order to prevent the hidden layer supervision module from overlearning the hidden layer features and falling into the local optimal solution, a weighted loss function L_{face} is designed:

$$L_{face}(W, w) = \sum_{m=1}^{M} \alpha_m \beta_m \left(W, w^{(m)}\right), \tag{12}$$

where W is the VGG16 parameter set, the weight of each supervision module is $w = (w^{(1)}, \cdots, w^{(M)})$, and $M = 3$. β_m represents the loss function of the mth supervision module; α_m represents the importance of the losses on different supervision modules. The loss weight of the deep supervision modules is increased, so that the output layer can learn

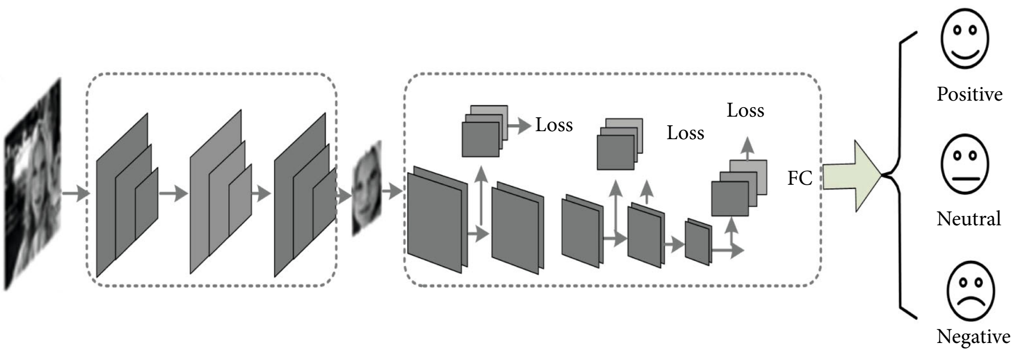


Figure 4: Face target sentiment analysis network.

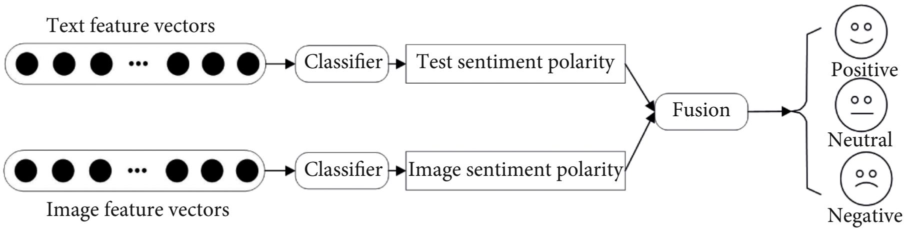


Figure 5: Text-image sentiment fusion module.

a feature map that is more conducive to facial sentiment recognition.

3.2.3. Global-Facial Image Sentiment Fusion Strategy. In order to comprehensively consider the face and the global image sentiment recognition model to determine the sentiment polarity of the image, a maximum fusion strategy is designed. The sentiment values predicted by the two models are recorded as $\boldsymbol{Y}_{face}$ and $\boldsymbol{Y}_{global}$, respectively:

$$\boldsymbol{Y} = \max\left(\boldsymbol{Y}_{face}, \boldsymbol{Y}_{global}\right) = \left(y_{pos}^{\max}, y_{neu}^{\max}, y_{neg}^{\max}\right), \tag{13}$$

where the three-dimensional vector $(y_{pos}, y_{neu}, y_{neg})$ represents the probability of positive, neutral, and negative sentiments, respectively.

3.3. Image-Text Fusion Module. Based on the text sentiment classification model and image sentiment classification model, a decision-level fusion method is designed for image-text fusion sentiment classification. The fusion module is shown in Figure 5.

The decision layer fusion is carried out by means of weighted summation. For each multimodal data sample s in the dataset, in the case of M kinds of modalities and C kinds of emotion classification, the obtained set of sentiment polarity probability is denoted by $\{p_{mc}(s), m = 1, 2, 3, \cdots, M\,; c = 1, 2, 3, \cdots, C\}$.

Since only two modalities of pictures and texts and three sentiments of positive, negative, and neutral are studied, $M = 2$ and $C = 3$, so the size of the sentiment polarity probability set is 6. The criterion of weighted summation is to weight and sum the C-type sentiment probability values of M kinds of modalities for each emotion classification c. A weighting coefficient is set for each modality, and the weighted summation can be expressed as

$$\begin{gathered} P_c(s) = \sum_{m=1}^{M} \lambda_m p_{mc}(s), \\ \sum_{m=1}^{M} \lambda_m = 1, \end{gathered} \tag{14}$$

where λ_m is the weighting coefficient of the mth modality. From this calculation, a new probability set $\{p_c(s), c = 1, 2, 3, \cdots, C\}$ is obtained, and the sentiment category corresponding to the $p_c(s)$ with the largest probability value is the classification result after data fusion.

4. Experiment

4.1. Experimental Datasets. The datasets used in the experiment include tourism reviews of image-text data collected from Chinese internet platforms Ctrip, Qunar, and Sina Weibo, called C-Tourism, and tourism review of image-text data collected on Twitter, called T-Tourism. The reviews are divided into three categories: positive, negative, and neutral. After preprocessing and manual annotation, the statistics of the final image-text tourism review datasets are shown in Table 1. The experimental environment is the Core i5 3.2 GHz CPU, 16 GB memory, and Windows 10 64-bit system, and Keras is used as the deep learning framework.

Table 1: Experimental datasets.

Dataset	Positive	Negative	Neutral	Total
C-tourism	4589	4715	2855	12159
T-tourism	5146	4658	1853	11657

4.2. Parameter Settings. The parameter settings for image sentiment analysis module are described in Sections 3.1 and 3.2. For the text sentiment analysis network, ten-fold cross-validation and grid search is used to verify different parameter combinations, and the best hyperparameter values are selected according to the evaluation results, as shown in Table 2.

4.3. Evaluation Metrics. In order to evaluate the effect of the classification model proposed in this paper, this paper adopts the commonly used rating indicators in the field of sentiment classification—accuracy (A), precision (P), recall rate (R), and the F1-socre (F1) to verity the tourism review sentiment classification performance.

The accuracy rate is the ratio of the number of correctly classified samples to the total number of samples, which is calculated as follows:

$$A = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \quad (15)$$

Here, TP is the number of positive samples correctly identified, TN is the number of negative samples correctly identified, FP is the number of positive samples incorrectly identified, and FN is the number of negative samples incorrectly identified.

The precision rate reflects how accurate the classification results are:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}}. \quad (16)$$

The recall rate reflects the completeness of the classification results:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (17)$$

The F1-score is the comprehensive indicator of recall rate and precision, as well as the degree of bias towards them:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \quad (18)$$

4.4. Comparison Models. The proposed method and the following comparison models are tested on the datasets, respectively, in order to verify the effectiveness and robustness of the proposed model:

(1) MCNN model [19] (multichannel convolution neural network). This model uses a multichannel CNN to classify the sentiment of tweet text, which is one of the earliest models that use deep learning models for sentiment analysis

(2) EMCNN model [20] (emoticon-semantic-enhanced multichannel convolution neural network). This model combines emojis to construct a semantic feature representation matrix and uses a multichannel CNN for feature learning to accomplish sentiment classification

(3) Bi-GRU model [21]. Bi-GRU uses traditional Glove word embedding and bidirectional GRU network to classify text sentiment

Table 2: Hyperparameter settings.

Parameters	Values
Activation function of the middle layer	Relu
Activation function of the last layer	Sigmoid
Optimizer	Adadelta
Convolution window size h	3, 4, 5
Number of filters m	100
Maximum number of iterations	80
Dropout rate	0.5

Table 3: Results on the C-Tourism dataset.

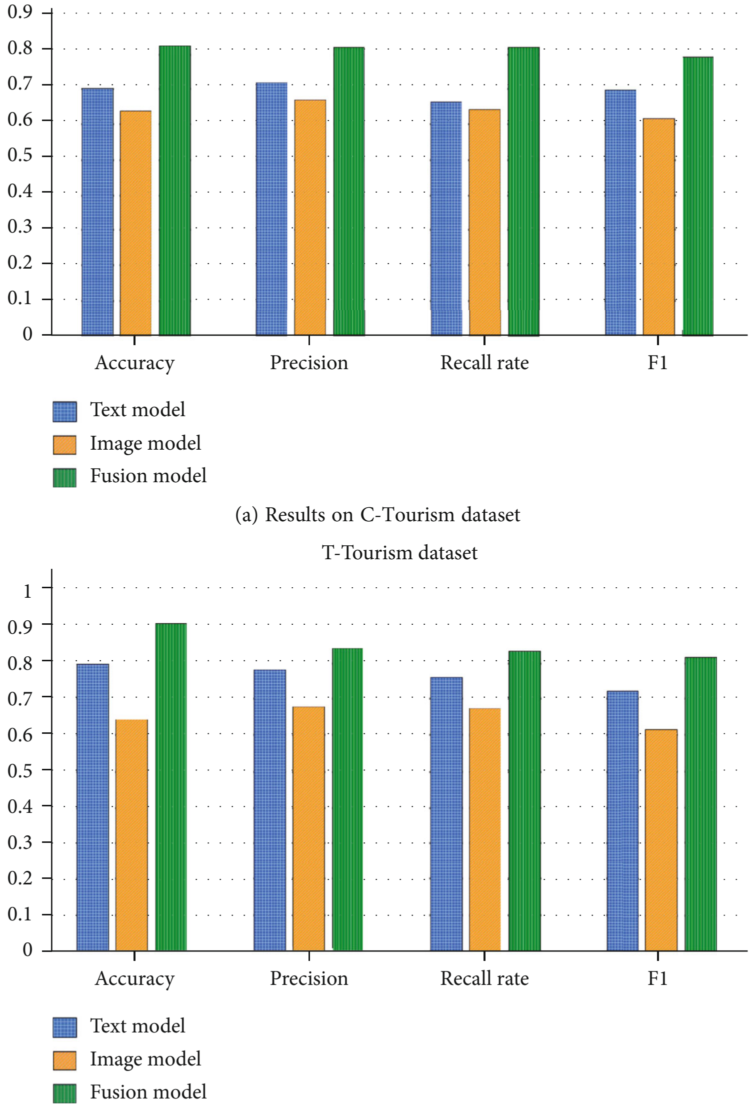
Models	A	P	R	F1
MCNN	0.5950	0.5620	0.4886	0.4812
EMCNN	0.6023	0.5622	0.5295	0.5389
Bi-GRU	0.6963	0.6891	0.6621	0.6378
EarlyFusion	0.7595	0.7717	0.7811	0.7401
Proposed method	0.8071	0.8211	0.8198	0.7936

Table 4: Results on the T-Tourism dataset.

Models	A	P	R	F1
MCNN	0.7057	0.7672	0.6841	0. 6499
EMCNN	0.7105	0.7530	0.6924	0.7077
Bi-GRU	0.7824	0.7629	0.7699	0.7210
EarlyFusion	0. 8527	0. 8177	0. 8192	0. 8011
Proposed method	0.9127	0.8477	0.8398	0.8236

Table 5: Binary classification results.

Models	Accuracy	
	C-Tourism	E-Tourism
MCNN	0.7562	0.7620
EMCNN	0.7693	0.7815
Bi-GRU	0.8054	0.8345
EarlyFusion	0.8539	0.9014
Proposed method	0.8890	0.9501



(a) Results on C-Tourism dataset

(b) Results on T-Tourism dataset

Figure 6: Ablation experiment results.

(4) EarlyFusion [22]. The model concatenates the image features obtained by the VGG13 model and the text features obtained by the Bi-GRU model and inputs them into the multilayer perceptron for sentiment classification

4.5. Results and Discussions. The sentiment classification results of each model on the C-Tourism dataset are shown in Table 3. It can be found that MCNN relies only on tourism review text for sentiment classification and has the worst classification performance. EMCNN slightly improves the sentiment classification performance by considering emojis in text sentiment classification. The Bi-GRU model considers the mining of the underlying semantics of text and outperforms MCNN and EMCNN. EarlyFusion and the proposed method combine text features and image features for tourism review sentiment classification and outperforms the text-dependent methods, proving the effectiveness of combining image and text features in sentiment analysis tasks. The proposed method achieves optimal performance because the proposed method combines four features in text classification with tourism review sentences and incorporates face detection and global image sentiment analysis network in image classification and further improves the performance of tourism review classification by fusion of image and text sentiment classification results.

The experimental results on the T-Tourism dataset are given in Table 4. Obviously, the classification effect of each model on T-Tourism dataset is better than that of the C-Tourism dataset, which indicates that English text is easier to perform sentiment classification than Chinese text. In addition, the collected T-Tourism dataset does not contain emojis. Since emojis are included in sentences in the form of words, when the short text does not contain emojis, the proposed model just does not extract the emojis and there is no need to modify the model, which verifies the flexibility of the proposed model and facilitates its usage in practical scenarios.

In order to further verify the sentiment classification performance of the proposed model, only positive and negative samples in the experimental dataset are retained, and the experimental results with the five models are shown in Table 5. The accuracy is used as the evaluation indicator in this experiment. It can be seen from Table 5 that after removing the neutral samples, the accuracy rate of all five models has improved significantly, and it is also verified that the English dataset still outperforms the Chinese dataset in the binary classification task. The proposed method achieves the best result of 95.01% on the English dataset. The results prove that neutral samples have greater recognition difficulty than positive and negative samples in the sentiment classification tasks.

4.6. Ablation Study. The role of each module in the proposed method is analyzed through ablation experiments. The experimental results are given in Figure 6. Among them, the image model uses only the image sentiment classification module to classify tourist review sentiment, which has the worst performance and proves the poor stability of determining tourist sentiment from images only. The text model uses only the text sentiment analysis network with multifeature fusion mechanism for sentiment classification. The fusion model achieves the best performance, which fully demonstrates that there is complementarity between images and text, and the combination of both can achieve optimal performance.

5. Conclusion and Future Works

In this paper, a multifeature fusion approach for sentiment analysis of tourism reviews containing images and texts is proposed. In the text sentiment classification module, word features, POS features, emoji features, and tag features are combined with tourism review sentences to enhance the ability of the proposed model to capture sentiment semantics. In the image sentiment classification module, a global and face-based image sentiment classification method is designed, and the results of the text module and image module are combined in the fusion module. The experimental results show that the fusion of the two modules achieves better performance at the decision level, indicating that the image-text feature fusion gives full play to the complementary roles of images and text and further improves the sentiment classification performance of tourism reviews.

However, in the process of facial emotional feature extraction, more complex situations are not considered in this paper, such as side face or occlusion. Therefore, future research will consider more complex and ubiquitous picture models. In addition, we will also study other more advanced classifiers to further improve the accuracy of emotional analysis of tourism comments.

Data Availability

All data included in this study are available upon request by contact with the corresponding author.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

[1] Y. Li, C. Hu, C. Huang, and L. Duan, "The concept of smart tourism in the context of tourism information services," *Tourism Management*, vol. 58, pp. 293–300, 2017.

[2] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, no. 2, pp. 617–663, 2019.

[3] J. Li, L. Xu, L. Tang, S. Wang, and L. Li, "Big data in tourism research: a literature review," *Tourism Management*, vol. 68, pp. 301–323, 2018.

[4] R. A. Hamid, A. S. Albahri, J. K. Alwan et al., "How smart is e-tourism? A systematic review of smart tourism recommendation system applying data management," *Computer Science Review*, vol. 39, article 100337, 2021.

[5] K. Lei, C. Wen, and X. Wang, "Research on the coordinated development of tourism economy based on embedded dynamic data," *Microprocessors and Microsystems*, vol. 82, article 103933, 2021.

[6] X. Zhou, M. Wang, and D. Li, "From stay to play - a travel planning tool based on crowdsourcing user- generated contents," *Applied Geography*, vol. 78, pp. 1–11, 2017.

[7] T. Chen, H. M. Salaheldeen, X. He, M. Y. Kan, and D. Lu, "Velda: relating an image tweet's text and images," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, 2015.

[8] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, https://arxiv.org/abs/1404.2188.

[9] X. Wang, Y. Liu, C. J. Sun, and B. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1343–1353, China, 2015.

[10] D. T. Vo and Y. Zhang, "Target-dependent twitter sentiment classification with rich automatic features," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015.

[11] V. Campos, B. Jou, and X. Giro-i-Nieto, "From pixels to sentiment: fine-tuning CNNs for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

[12] M. Li, W. Li, F. Wang, X. Jia, and G. Rui, "Applying BERT to analyze investor sentiment in stock market," *Neural Computing and Applications*, vol. 33, no. 10, pp. 4663–4676, 2021.

[13] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: when attention meets tree-structured recursive neural networks," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1008–1017, New York, 2016.

[14] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Systems*, vol. 161, pp. 124–133, 2018.

[15] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, “Image-text sentiment analysis via deep multimodal attentive fusion,” *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.

[16] X. Fu, W. Liu, Y. Xu, and L. Cui, “Combine HowNet lexicon to train phrase recursive autoencoder for sentence- level sentiment analysis,” *Neurocomputing*, vol. 241, pp. 18–27, 2017.

[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[18] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, “Learning supervised scoring ensemble for emotion recognition in the wild,” in *Proceedings of the 19th ACM international conference on multimodal interaction*, pp. 553–560, New York, 2017.

[19] D. H. Pham and A. C. Le, “Exploiting multiple word embeddings and one-hot character vectors for aspect- based sentiment analysis,” *International Journal of Approximate Reasoning*, vol. 103, pp. 1–10, 2018.

[20] Y. Tao, X. Zhang, L. Shi, L. Wei, Z. Hai, and J. A. Wahid, “Joint embedding of emoticons and labels based on CNN for microblog sentiment analysis,” in *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pp. 168–175, IEEE, 2019.

[21] S. Sachin, A. Tripathi, N. Mahajan, S. Aggarwal, and P. Nagrath, “Sentiment analysis using gated recurrent neural networks,” *SN Computer Science*, vol. 1, no. 2, pp. 1–13, 2020.

[22] Z. Zhao, H. Zhu, Z. Xue et al., “An image-text consistency driven multimodal sentiment analysis approach for social media,” *Information Processing & Management*, vol. 56, no. 6, pp. 102097–102219, 2019.