International Journal of Information Systems in the Service Sector
Volume 15 • Issue 1 • January-December 2024

Enhancing Multimodal Tourism Review Sentiment Analysis Through Advanced Feature Association Techniques

Peng Chen
Sanya Institute of Technology, China

Lingmei Fu
Hainan Provincial Sports Academy, China

ABSTRACT

The development of tourism services presents significant opportunities for extracting and analyzing customer sentiment. However, with the advent of multimodality, travel reviews have brought new challenges. Early methods for detecting such reviews merely combined text and image features, resulting in poor feature correlation. To address this issue, our study proposes a novel multimodal tourism review sentiment analysis method enhanced by relevant features. Initially, we employ a fusion model that combines BERT and Text-CNN for text feature extraction. This approach strengthens semantic relationships and filters noise effectively. Subsequently, we utilize ResNet-51 for image feature extraction, leveraging its ability to learn complex visual representations. Additionally, integrating an attention mechanism further enhances modality correlation, thereby improving fusion effectiveness. On the Multi-ZOL dataset, our method achieves an accuracy of 90.7% and an F1 score of 90.8%. Similarly, on the Ctrip dataset, it attains an accuracy of 83.6% and an F1 score of 84.1%.

KEYWORDS

Attention Mechanism, Enhanced Associative Features, Feature Synthesis, Multimodal Sentiment Analysis, Tourism

As tourism experiences gain more importance, understanding the correlation between tourism evaluations and actual experiences has become crucial. Existing research focuses primarily on single-modal data, often neglecting non-textual information such as images and ratings. To address this gap, it is essential to explore the role of multimodal data in tourism reviews for more comprehensive sentiment analysis.

Tourists' opinions are multimodal, including text, images, and ratings, posing significant challenges for sentiment analysis. Current research on integrating image information in tourism reviews is limited and needs further exploration. By combining text and image data, we can achieve a better understanding of tourists' evaluations, providing valuable feedback for the tourism industry.

Some models face challenges in fully capturing semantic intricacies. The evolution of online tourism feedback toward multimodal formats necessitates new approaches for analyzing complex emotional cues. Therefore, integrating correlated features within multimodal datasets is imperative for precise sentiment analysis.

This research introduces an advanced algorithm for multimodal sentiment analysis of online travel reviews. Our combined approach optimizes feature extraction, improves predictive precision, and enhances the correlation between modalities, resulting in a rich, multimodal dataset for sentiment

DOI: 10.4018/IJISSS.349564

analysis. By leveraging interconnected data types, our method significantly enhances accuracy and stability in sentiment predictions.

In summary, this study aims to develop a robust multimodal sentiment analysis algorithm, offering deeper insights and more accurate sentiment predictions from tourism reviews. In this research, we have undertaken the following initiatives:

(1) To overcome the challenges posed by singular models in extracting text features, which often overlook intricate semantic connections because of their limited representation capabilities, this study introduces a hybrid text feature extraction approach. This method combines the BERT model, renowned for its robust semantic representation, with the Text-CNN model, acclaimed for its proficiency in identifying essential local attributes. The integration of these models enriches and refines the process of text feature extraction.
(2) Addressing the challenge of insufficient feature linkage stemming from disparities among various types of modal data, this work employs an attention mechanism within the framework. This addition to the ResNet-51 model allows for a more precise capture of image attributes that correlate directly with textual elements, thus strengthening the feature connectivity between modalities. This advancement significantly boosts the efficacy of sentiment analysis in multimodal tourism reviews.

RELATED WORK

Sentiment classification is crucial for understanding tourism reviews, focusing on subjective emotions (Chen et al., 2020; Krishnan et al., 2022; Momani et al., 2022). Traditional single-modal analysis misses important cues in text and images (Ye et al., 2022).

Single-Modality Sentiment Classification Methods

Sentiment classification is crucial for understanding tourism reviews, focusing on subjective emotions (Chen et al., 2020; Krishnan et al., 2022; Momani et al., 2022). Traditional single-modal analysis, which often examines only text or images, misses important emotional cues (Ye et al., 2022). This study introduces an advanced algorithm combining BERT and Text-CNN for text extraction, ResNet-51 for image extraction, and an attention mechanism to integrate multimodal data, enhancing sentiment prediction accuracy.

Single-modality sentiment analysis has been pivotal in analyzing text (Wang & Shin, 2019) and images (Rao et al., 2020), using statistical methods like term frequency and inverse document frequency (Puh & Bagić, 2023). Advances in pretrained models like BERT (Devlin et al., 2018), GPT-2 (Veyseh et al., 2021), and RoBERTa have improved text sentiment analysis by capturing complex language structures. BERT-CNN integrations (Abas et al., 2022) further enhance the capture of emotional nuances. Prompt learning (Liu et al., 2023) has facilitated few-shot learning and improved semantic comprehension.

Visual sentiment analysis initially relied on handcrafted features such as composition and texture (Machajdik & Hanbury, 2010) and concepts like balance and harmony (Zhao et al., 2014). Innovations like adjective noun pairs (ANPs; Borth et al., 2013) and their emotional implications (Li et al., 2018) have been significant. Recently, deep neural networks and attention mechanisms have improved visual sentiment analysis by focusing on emotionally significant image areas (Yang et al., 2021; You et al., 2017). Integrating visual and textual analyses into a multimodal approach promises enhanced precision (Yang et al., 2021; Zhang et al., 2022).

Despite progress, challenges remain, including a scarcity of annotated datasets and model efficiency issues. Further research is needed to refine these multimodal techniques and improve their practical applications.

Multimodal Sentiment Classification Methods

Contemporary sentiment analysis research focuses on integrating different modalities and their combinations in dual or triple formats (Al-Tameemi et al., 2022). Enhancements in multimodal sentiment analysis involve techniques for extracting features from individual channels (Ajitha et al., 2021; Zaw & Tandayya, 2022) and strategies for integrating these features (Chatterjee et al., 2023; Gandhi et al., 2023; Kaur & Kautish, 2022).

Early fusion methods, like the EF-LSTM introduced by Hou et al. (2022), combine features from various modalities early in the processing pipeline. However, this approach can introduce superfluous data. Conversely, late fusion techniques operate at the decision making stage but struggle with inter-modality dynamics. Recent models, such as the two-tower architecture used in tourism recommendations (Cui et al., 2024), leverage both textual and visual data to improve accuracy and user experience. Additionally, Wei et al. (2022) and Zhang et al. (2023) enhance sentiment analysis by integrating features from text and images.

To address the complexity and high dimensionality in multimodal data, the memory fusion network (MFN) by Zadeh et al. (2018) incorporates attention mechanisms and gated memory to trace modality interactions. The multimodal factorization model (MFM) by Tsai et al. (2018) factorizes data into shared and unique components for each modality, fostering a comprehensive understanding.

Attention mechanisms have significantly enhanced model capabilities by prioritizing key information, improving the integration of multimodal inputs (Rahman et al., 2022). The MAG-BERT technique adapts BERT to integrate visual and acoustic data alongside text, refining the model through fine-tuning. However, these cross-modal strategies often overlook the distinct attributes of each modality during single-modality feature extraction.

In summary, while multimodal approaches show promise, challenges remain in efficiently integrating diverse data types and fully capturing the unique characteristics of each modality. Further research is needed to refine these techniques for practical applications.

METHODOLOGY

Online travel review sentiment analysis is a binary classification task in which $D = \{N_1, N_2, \ldots, N_m\}$, a collection of multimedia travel reviews. $y \in \{0, 1\}$ is each travel review $N_i \in D$ label, where 0 belongs to negative sentiment and 1 belongs to positive sentiment. Each travel review N_i consists of related text information (T) and image information (I), denoted as $N_i = T_i \cup I_i$. Thus, a model is defined $F{:}D \rightarrow y$ that categorizes each review into a predefined class label $y \in \{0, 1\}$, shown in the Equation 1:

$$F(N) = \begin{cases} 0, \text{ if } N_i \text{ is fake} \\ 1, \text{ if } N_i \text{ is real} \end{cases} \quad (1)$$

where $F(N)$ is a function measuring sentiment in tourism reviews. Figure 1 depicts the framework of this algorithm:

Text Feature Extraction

In the travel review sentiment analysis task, textual data usually contains rich semantic information and complex relationships. Therefore, this proposal uses BERT to extract textual features of travel reviews and then uses a Text-CNN model to further extract deeper potential semantic information.

Figure 1. Framework of travel comment sentiment analysis algorithm based on associative feature enhancement

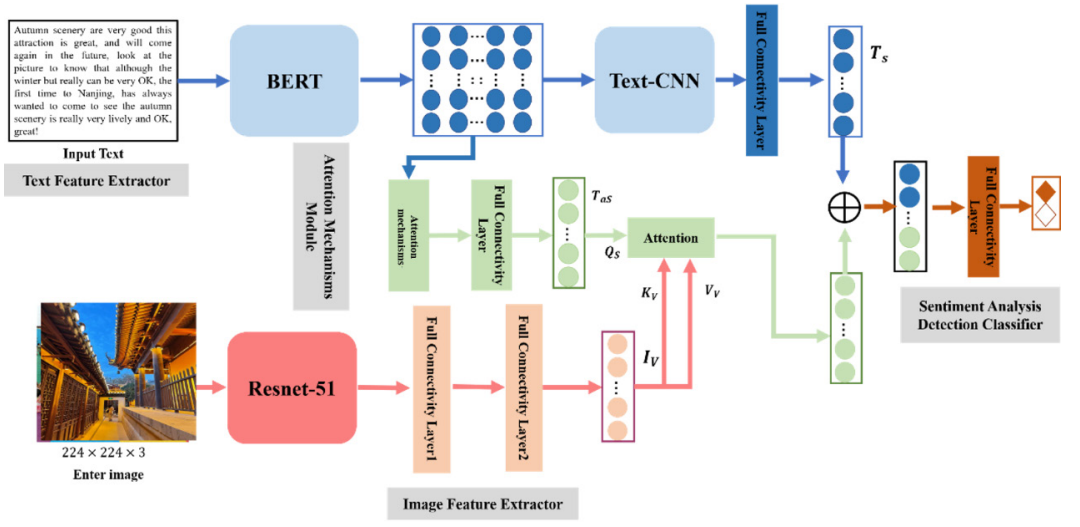
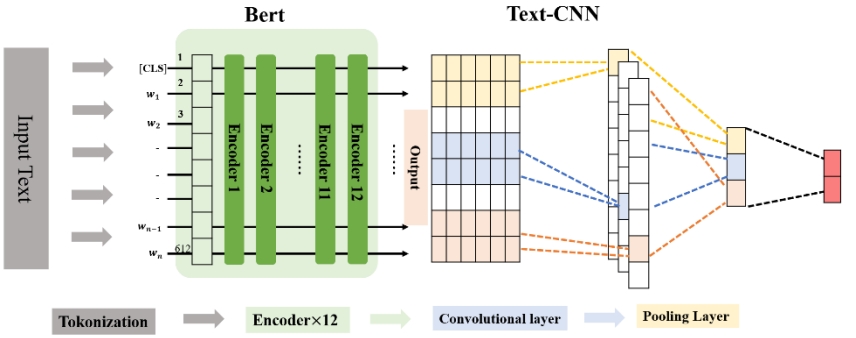


Figure 2. Text feature extractor framework



The BERT model captures potential semantic information from input words and sentences by labeling words with random masks while vectorially representing each masked word, and in restoring the words at the position of the mask, capturing potential semantic and textual features. Therefore, there is a need to enhance travel review text features by noise reduction of key information using Text-CNN models to improve the accuracy of travel review detection.

As shown in Figure 2, first, the text content is preprocessed and converted into embedding of word vectors, embedding of segment vectors, and embedding of position vectors, and at the same time, the three embedding vectors are summed up as the input sequence of BERT, through which BERT can learn the text corpus and extract features. The whole input sequence is passed layer by layer through the feed-forward network to the 12-layer transformer model, and the output is represented as T_b.

Text features of travel reviews are extracted from BERT T_b as input to the Text-CNN model, which consists of three main modules: a word embedding layer, a convolutional layer, and a maximum pooling layer. The convolutional layer makes use of the window size of the embedded text for the embedded text n of the convolution kernel to slide over the sentence and dot product with local word embeddings to extract local features. Maximum pooling performs dimensionality reduction on the

information features in the domain and enhances positional invariance. Specifically, the convolution operation is formulated in Equation 2:

$$t_i = \sigma\left(W_c \times T^b_{i:i+n-1}\right) \tag{2}$$

where $T_b \epsilon T^{l \times d}$, where d denotes the dimension of the hidden layer for BERT, and l denotes the length of the input text sequence. W_c denotes the convolution kernel weights, and i denotes the weight of the convolution kernel starting from the first word in the input sequence. i word in the input sequence $\sigma(\cdot)$ denotes the modified linear unit (Leaky ReLU) activation function. After the convolution kernel performs the convolution operation on each word in sequence, the extracted feature mapping is as shown in Equation 3:

$$t = [t_1, t_2, \ldots, t_{l-n+1}] \tag{3}$$

Where the feature vector $t \epsilon T^{l-n+1}$, which is then used as input to the maximum pooling layer, is computed to extract the maximum features, and the computational process can be expressed as in Equation 4:

$$\hat{t}^k = max(r) \tag{4}$$

The output feature vector after pooling computation can be expressed as in Equation 5:

$$\hat{t}^k = \hat{t}_1, \hat{t}_2, \ldots, \hat{t}_{j_n} \tag{5}$$

where j denotes the number of convolution kernels and the feature vector $\hat{t}^k \epsilon T_{j_n}$.

For the convolution operation, different convolution kernels can be selected to extract contextual information with different breadth, i.e., fine-grained feature sequence information. After processing all the features, the pooling results are spliced to obtain the feature vector $T_g \epsilon T^{g \times j}$ where g denotes g different kinds of window sizes, and $g \times j$ denotes that there are j different sizes of convolution kernels and g kinds of sizes of convolution kernels. In order to match the dimensional features of the image features uniformly, the text feature vector output from the pooling layer is passed to the employee fully connected layer for dimensionality proofreading, which can be denoted as in Equation 6:

$$T_S = \sigma\left(W_{sf} \times T_g\right) \tag{6}$$

where the final text feature vector $T_S \epsilon T^p$ and the W_{sf} is denoted as the weight of the fully connected layer.

Image Feature Extraction

Image feature extraction serves the same purpose as text feature extraction. This study chose ResNet-51 as the foundation for the image feature extraction module. In this section, the parameters of the ResNet-51 model were fixed while the feature vector dimensions and the parameters of the fully connected layers were simultaneously updated. Specifically, the travel review image data was preprocessed to be converted into the input of ResNet-51 model; then the ResNet-51 model extracted the image feature vectors, and finally the extracted vectors. First, an image information was extracted from a given travel review dataset I as the input of ResNet-51, which can be expressed as in Equation 7:

$$I_{V_{ResNet}} = ResNet(I) \tag{7}$$

Among them $I_V \epsilon I^p$, and the $I_{V_{ResNet}}$ denotes the image features extracted for the *ResNet* -51 image features extracted by the model. By passing the extracted image feature vector $I_{V_{ResNet}}$ to the fully connected layer, the final image feature vector can be expressed as in Equation 8:

$$I_v = \sigma\left(W_{vf2} \times \left(W_{vf1} \times I_{ResNet}\right)\right) \tag{8}$$

Of these, the W_{vf1} and W_{vf2} are the weight matrices of the two fully connected layers, respectively, and σ denotes the ReLU activation function.

Attention mechanisms

Fusing two modal features, text and image, is an effective method for sentiment analysis of travel reviews, since these two modalities contain different semantic information, which can complement and promote each other. However, most of the existing methods directly use spliced feature vectors for fusion, which leads to a sharp increase in dimensionality, thus increasing the computational complexity and memory occupation of the algorithms and decreasing the efficiency. In addition, since text and image features have different expressions and information volumes with great variability, direct splicing also leads to insufficient feature correlation between different modalities, causing some features to be masked or ignored, thus affecting the accuracy of the algorithm. Therefore, in this study, the attention mechanism was used to identify image features that are associated with the semantics of the text and assign higher weights to these image features.

This section uses the scaled dot product self-attention mechanism, which computes the attention score by matrix multiplication and can be implemented efficiently and with computational effectiveness using parallel computing. The scaled dot product self-attention mechanism better captures global dependencies owing to the ability to weight different input elements according to the relevance of the inputs. In addition, the mechanism can handle inputs of different lengths by computing the attention scores among all input elements, making it suitable for processing long sequences.

Compared with the text feature vectors further extracted by Text-CNN, the text feature vectors extracted on the basis of BERT can more adequately retain the rich contextual features, making them more conducive to finding the connection with image features. Therefore, the text features of the BERT-based model extracted comments are directly T_b passed to the fully connected layer with adjusted dimensions, and the output text features are represented as T_{aS}. Then the text feature vector T_{aS} mapped as a query vector Q_S and the image feature vector extracted according to the ResNet-51 model I_V is mapped as a value vector V_V and key vectors K_V where $T_{aS} \epsilon T^p$ and $I_V \epsilon T^p$. The query vector is first computed by dot product Q_S and the key vector K_V. The attention score between the query vector and the key vector is then scaled with the square root of the dimension and the Softmax activation function is applied to compute the attention weights, which are finally combined with the value vector V_V. The weighting calculation outputs the image feature vector I_{∇}. The formula is shown in Equations 9 and 10:

$$\text{Attention}\left(Q_S, K_V, V_V\right) = \text{softmax}\left(\frac{Q_S K_V^T}{\sqrt{p}}\right) V_V \tag{9}$$

$$I_{\nabla} = \text{Attention}\left(Q_S, K_V, V_V\right) \tag{10}$$

Where the query vector Q_S, the value vector V_V, and the key vector K_V all have dimensions of p.

After computation based on the BERT model and the Text-CNN model, the text feature vector is obtained T_S. After computation based on the ResNet-51 model and the attention mechanism, the image feature vector is obtained $I_{\tilde{V}}$. In this section, the obtained feature vectors of two modalities are spliced to obtain a multimodal feature vector with high information content that can be expressed as in Equation 11:

$$R_c = T_S \oplus I_{\tilde{V}} \tag{11}$$

Sentiment Classifier

A sentiment classifier can automatically categorize the sentiment of travel review texts. A multimodal feature vector obtained R_c as the input to this model, passed into a fully connected layer based on the Softmax activation function, and the output is a probability distribution. Therefore, it is used in this section to calculate the probability that the input review content is negative sentiment, which can be expressed as in Equation 12:

$$p\left(R_c^i\right) = \text{Softmax}\left(W_f \cdot R_c^i\right) \tag{12}$$

where the multimodal feature vector R_c dimension is the sum of the text feature vector dimension and the image feature dimension, i.e. $R_c \in R^{2p}$ .i is denoted as the dimension of the i multimodal corpus of travel reviews, and W_f is the weight parameter of the fully connected layer.

In this section, the cross-entropy function is chosen to calculate the logarithmic difference between the forward labeling vector and the model's predicted output vector to measure the similarity, which can be expressed as in Equation 13:

$$\mathscr{L}_{loss}(\theta) = -E_{(c,y)\sim(C,Y)}\left[ylogp(R_c) + (1-y)\log\left(1-p(R_c)\right)\right] \tag{13}$$

When training a model, our goal is to minimize the cross-entropy loss by adjusting the parameters denotes θ the model parameters, as shown in Equation 14:

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}_{loss}(\theta) \tag{14}$$

During model training, parameters are adjusted using the stochastic gradient descent algorithm based on the loss function's gradient direction. In addition, the stochastic gradient descent algorithm needs to choose a suitable learning rate to control the step size of the parameter update. Therefore, Adam optimizer using adaptive learning rate is chosen to optimize the classifier for travel review sentiment.

EXPERIMENTATION AND ANALYSIS

In this section, two datasets for sentiment detection tasks in reviews are first introduced. Subsequently, the performance of the proposed multimodal tourism review detection algorithm based on attention mechanisms is evaluated; finally the experimental results are analyzed.

Experimental Data Set and Evaluation Indicators

Multi-ZOL dataset

The Multi-ZOL dataset collects user comments about mobile phones, which are extracted from IT-related information and the business portal website ZOL.com, and are labeled and filtered. This

Table 1. Statistical information of multi-ZOL dataset

Causality	statisticians
Number of comments	5228
Label count	10
Average number of words per comment	315.11
Maximum number of words per comment	8511
Minimum number of words per comment	5
Average number of images per comment	4.5
Maximum number of images per comment	111
Minimum number of images per comment	1

Table 2. Sample multi-ZOL dataset

Example	Text	Image	Sentiment label
1	The phone is quite good, black is quite cool, the display is good, there is a full eye protection mode, beautiful appearance, running smoothly, is not quite adapt to the input method, the battery one day a charge certainly want, is to buy in the official flagship store in Taobao did not send headphones is quite unhappy, psychological imbalance! Battery capacity is small, on the shortcomings of this.		Negative
2	(1)The appearance of low-key, calm, introverted, business sense, loved by men: (2)Split-screen, display, color performance is still very good at this price point;. (3)Provide posing posture guidance, support for front soft light;		Positive

paper only selects graphic-text comments from the original data, which contain both text and images, totaling 5,288. Each graphic-text datum includes a text content, an image set, and a sentiment score ranging from 1 to 10. Table 1 and Table 2 display the detailed statistics of this dataset.

Ctrip Dataset

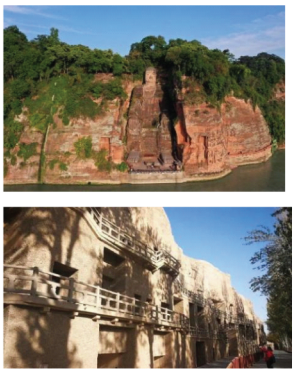
Currently, research on sentiment analysis of tourism reviews is still in its early stages, and the lack of corpus resources is a significant challenge facing this field of study. Therefore, in order to achieve sentiment analysis of reviews in the tourism domain, it was necessary to collect and annotate relevant data to improve the model's understanding and classification of review sentiments. In this paper, by constructing the Ctrip Tourism Review Sentiment Analysis Dataset, valuable data resources are provided for research in this field. This study focused on constructing a tourism review dataset, which includes 1,243 tourist attractions and 37,518 graphic-text reviews. However, since these data are collected from travel websites using data crawling, they do not possess standardized structured text features. Therefore, researchers first conducted necessary data cleaning work. Table 3 and Table 4 display the detailed statistics of this dataset.

This dataset and the collated annotation information will be useful for conducting more in-depth multimodal sentiment analysis studies to better understand and analyze the sentiment information in

Table 3. Statistical information of Ctrip dataset

Causality	Statistics
Number of comments	37518
Label count	5
Average number of words per comment	189.2
Maximum number of words per comment	5712
Minimum number of words per comment	4
Average number of images per comment	4
Maximum number of images per comment	19
Minimum number of images per comment	1

Table 4. Sample Ctrip data set

Example	Text	Image	Sentiment label
1	Early holiday, drive to the scenic area after arriving by the ticket hall building into the scenic area, the elderly over 60 years old free tickets. We started from the 5th cave tour, because there are accompanied by the elderly children, we entered the scenic area after the choice of sightseeing bus, round trip 15 yuan, the price is still very friendly. Through the introduction of the guide to the various caves of the Buddha, understand the history of the Northern Wei Dynasty Xianbei culture and Han culture mingling		Positive

multimodal data. This cleansing and organizing of the data is essential for building effective models for sentiment analysis.

Experimental Details

In the text feature extraction module, the BERT-base model processed the Multi-ZOL dataset, while BERT-base-Chinese was utilized to handle the Ctrip dataset. In the Text-CNN, there were 20 convolutional kernels, with four different window sizes (1–4). To prevent overfitting, the parameters of the pretrained VGG19 and BERT networks were kept frozen during training. The hidden vector dimension of the last fully connected layer was set to 32; thus the dimension of the extracted feature vectors was also 32.

In the image feature extraction module, preprocessed images (226×226×3) were input to the ResNet-51 model, yielding feature vectors of size 1,000. Subsequently, the dimensions of the last two fully connected layers of the model were set to 512 and 32, respectively. Finally, the extracted vectors were sequentially passed through the two fully connected layers, resulting in image feature vectors of dimension 32.

Comparative Tests

The previous section outlined the experiment details. To evaluate the proposed algorithm, several state-of-the-art methods were chosen as baseline approaches. First, we introduced these baseline methods; then we compared the experimental outcomes with them.

Text

The Text-CNN method is utilized to extract text features from travel reviews (Zhang et al., 2020).

Vision

Image features are extracted with a pre-trained VGG-19 network and then processed via a fully connected layer using the ReLU function to analyze the sentiment of travel reviews, producing an output size of 32 (Zhang et al., 2020).

VQA

This method produces textual responses that correspond to a provided image (Antol et al., 2015). To ensure equitable comparison of experimental outcomes, the ultimate multi-classification output layer is adapted to a binary classification output layer.

ATT-RNN

An attention mechanism is employed to integrate text, image features, and social context features (Jin et al., 2017). For equitable comparison, we excluded the social context feature extraction component while maintaining the other configurations unchanged in the experiments.

Neural Talk

This method produces image descriptions through a deep recursive framework (Vinyals et al., 2015). The feature representation is obtained by averaging RNN outputs at each time step, then inputting this into a fully connected layer with a dimension of 32, and finally predicting outcomes using cross-entropy loss.

EANN

In the analysis of multimodal travel reviews, an end-to-end adversarial event network is employed, featuring a multimodal feature extractor and a sentiment detector for travel reviews, while the event distinguisher is omitted to ensure fair comparisons (Wang et al., 2018).

MVAE

In this analysis of multimodal travel reviews, we utilized an adversarial event network possessing a feature extractor and sentiment detector, omitting the event distinguisher to ensure equitable comparisons (Khattar et al., 2019).

SAFE

This method investigates the influence of cross-modal similarity by introducing an auxiliary goal function designed to measure the discrepancy between similarity scores and labels (Zhou et al., 2020).

SpotFake

This sentiment evaluator for multimodal journey assessments, utilizing BERT for textual attributes and VGG-19 for visual attributes, combines their vectors and employs Softmax for sentiment assessment (Singhal et al., 2020).

Table 5. Comparison of experimental results between multi-ZOL dataset and Ctrip

	Multi-ZOL				Ctrip (Chinese company)			
Methods	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
Text (Zhang et al., 2020)	0.763	0.827	0.683	0.748	0.562	0.598	0.541	0.568
Vision (Zhang et al., 2020)	0.615	0.615	0.677	0.645	0.546	0.695	0.518	0.593
VQA (Antol et al., 2015)	0.773	0.780	0.782	0.781	0.631	0.765	0.509	0.611
NeuralTalk (Jin et al., 2017)	0.717	0.683	0.843	0.754	0.610	0.728	0.504	0.595
ATT-RNN (Vinyals et al., 2015)	0.779	0.778	0.799	0.789	0.664	0.749	0.615	0.676
EANN (Wang et al., 2018)	0.827	0.847	0.812	0.829	0.715	0.822	0.638	0.719
MVAE (Khattar et al., 2019)	0.824	0.854	0.769	0.809	0.745	0.801	0.719	0.758
SAFE (Zhou et al., 2020)	0.839	0.824	0.840	0.832	0.760	0.826	0.731	0.775
SpotFake (Singhal et al., 2020)	0.821	0.848	0.825	0.847	0.731	0.752	0.703	0.719
P-tuning (Liu et al., 2022)	0.851	0.863	0.851	0.862	0.755	0768	0.834	0.824
Input-tuning (An et al., 2022)	0.883	0.898	0.872	0.885	0.778	0.773	0.814	0.791
MBRT	**0.907**	**0.908**	**0.907**	**0.908**	**0.836**	**0.831**	**0.868**	**0.841**

P-Tuning

P-tuning is a method for enhancing the performance and adaptability of PLMs by iteratively fine-tuning parameters on specific target tasks (Liu et al., 2022).

Input-Tuning

Input-tuning introduces a superior method to adapt new inputs to frozen PLMs by fine-tuning continuous prompts and input representations, achieving better results than prompt-tuning in natural language generation (NLG) tasks (An et al., 2022).

Table 5 shows that the proposed method significantly surpasses nine other methods in terms of accuracy, precision, recall, and F1 score across both the Multi-ZOL and Ctrip datasets. The proposed method achieves accuracies of 90.7% on the Multi-ZOL dataset and 83.6% on the Ctrip dataset. The superior performance on the Multi-ZOL dataset is due to its more balanced distribution of events and higher quality and quantity of images compared to the Ctrip dataset, which contains over 10,000 reviews but only 514 images.

The results show that the unimodal text approach performs better in terms of accuracy, while the unimodal image approach excels in precision on the Ctrip dataset. This suggests that text and image provide different recognition features in sentiment analysis. Aggregating these unimodal features improves overall performance. Additionally, the unimodal text approach outperforms the unimodal image approach in all metrics on the Multi-ZOL dataset, indicating that text data contains rich semantic information crucial for accurate sentiment analysis.

The importance of image information in travel review sentiment analysis should not be underestimated. Combining image information with text provides richer features and improves accuracy. As shown in Figures 3 and 4, multimodal methods, starting from VQA, outperform unimodal methods on both datasets, highlighting the advantages of utilizing multimodal information.

However, not all multimodal methods perform equally well. VQA and Neural Talk perform poorly overall, while ATT-RNN shows better results, demonstrating the effectiveness of attention mechanisms in enhancing sentiment detection. EANN and MVAE have limitations in feature extraction and fusion,

Figure 3. Comparison of accuracy and F1 value of the proposed method with other methods under Ctrip dataset

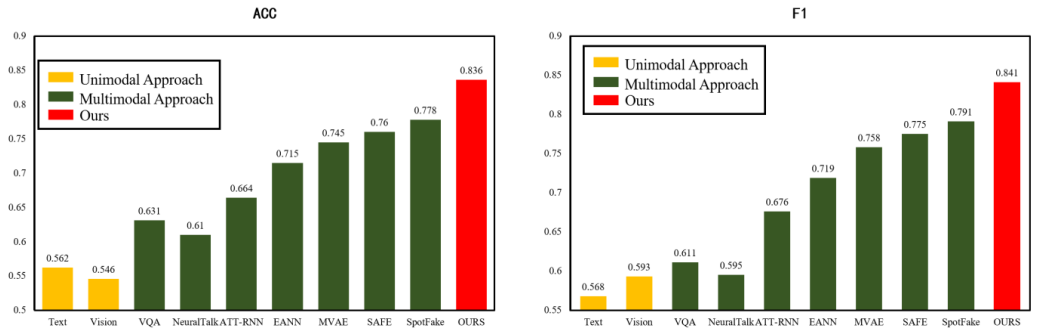
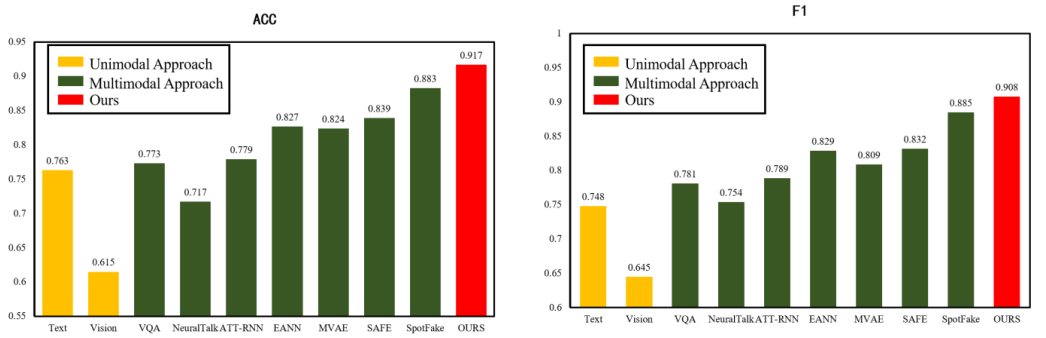


Figure 4. Comparison of accuracy and F1 value of the method with other methods under multi-ZOL dataset



leading to insufficient feature correlation and mediocre performance. EANN's multimodal feature extraction through an event adversarial network lacks careful processing in text and image fusion, and MVAE's bimodal variational autoencoder fails to provide sufficient feature fusion, limiting its practical application.

The SAFE method performs poorly on the Ctrip dataset owing to the low number and quality of images, which affects its image feature extraction and fusion capabilities. SpotFake utilizes BERT to extract and fuse textual and image features. However, it does not optimize the extracted unimodal features, leading to insufficient feature correlation and masking important features.

Therefore, the proposed method enhances both text and image feature extraction using Text-CNN for deeper latent semantic information and ResNet-51 for more discriminative image features. Additionally, the attention mechanism strengthens the link between text and image features, leading to enhanced performance in multimodal sentiment analysis.

Compared to SpotFake, the proposed method shows significant innovation and optimization in feature extraction and fusion, resulting in 5.8% and 2.4% higher accuracy. In summary, the proposed algorithm effectively improves travel review sentiment analysis through its optimized and innovative approach to multimodal data.

Ablation Experiments

To evaluate the impact of each component in the proposed algorithm, ablation experiments were carried out on the Multi-ZOL dataset. By systematically removing key components of the algorithm,

Table 6. Ablation experiments

Methodologies	Modal (Computing, Linguistics)	Accuracy	Accuracy	Recall Rate	F1 Value
Text-CNN	unimodal text	0.763	0.827	0.683	0.748
BERT	unimodal text	0.836	0.849	0.822	0.839
BERT+Text-CNN	unimodal text	0.861	0.892	0832	0.864
ResNet-51	unimodal images	0.623	0.621	0.701	0.650
-TC	multimodal	0.827	0.872	0.774	0.828
-RN	multimodal	0.876	0.871	0.852	0.880
-ATT	multimodal	0.881	0.903	0.862	0.893
MBRT	multimodal	0.907	0.908	0.907	0.908

we observed the changes in model performance, allowing us to evaluate the impact of each module. The ablation experiment setup was as follows:

Text-CNN: Sentiment analysis using only text content. Pre-trained Word2Vec vectors replaced BERT-generated vectors, and Text-CNN was used for feature extraction.

BERT: Sentiment analysis using only text content. BERT extracted text features, which were passed to a fully connected layer with Softmax for prediction.

BERT + Text-CNN: Only text content was used by combining BERT and Text-CNN for feature extraction, with no image features involved.

ResNet-51: Only image content was used. ResNet-51 extracted image features, which were reduced in dimensionality and passed to a binary classifier.

Text-CNN: Text features were extracted using BERT, but not further processed by Text-CNN.

ResNet: ResNet-51 was replaced with VGG-19 for image feature extraction, while text features from BERT were retained.

ATT: Text and image features were extracted using BERT and ResNet-51, respectively, but the attention mechanism was not used for feature fusion. Features were directly combined without attention.

These tests determine the roles of text feature extraction (BERT, Text-CNN), image feature extraction (ResNet-51, VGG-19), and the attention mechanism in enhancing the overall performance of the algorithm.

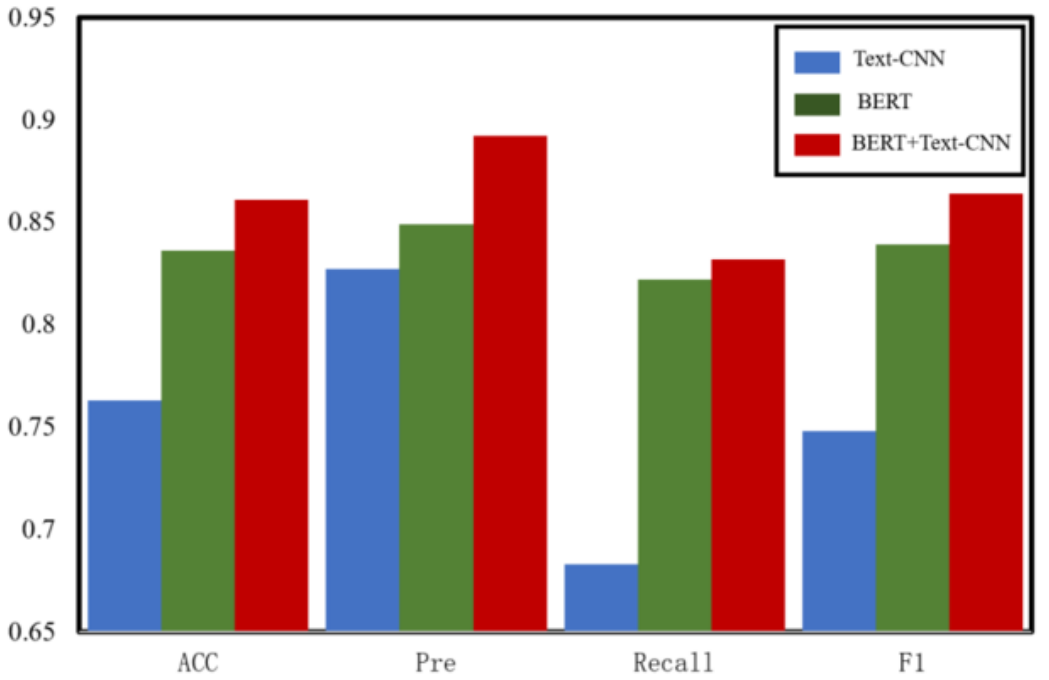
Table 6 presents the accuracy, precision, recall, and F1 scores for each module in the ablation experiments on the Multi-ZOL dataset. The results highlight that the unimodal text method consistently outperforms the unimodal image method, indicating that textual information is more discriminative for travel review sentiment analysis.

Among the text feature extraction methods, the BERT+Text-CNN model showed the best performance, as seen in Figure 5. The BERT model improved accuracy by 9.5% over Text-CNN alone due to its superior ability to capture the semantics and context of long text sequences. The fusion of BERT and Text-CNN leverages both models' strengths, combining sensitivity to local features with contextual understanding.

Multimodal methods generally outperform unimodal approaches, as shown in Figure 6. Removing the Text-CNN module resulted in a decrease in all evaluation metrics, underscoring its importance in extracting text features. While BERT demonstrated strong text feature extraction, Text-CNN complemented it by capturing local feature information, enhancing overall performance.

These findings confirm that utilizing both BERT and Text-CNN for text feature extraction significantly improves sentiment analysis in multimodal scenarios. The experiments demonstrate that

Figure 5. Comparison of unimodal text methods



effective multimodal sentiment analysis requires integrating various types of text feature extraction models to maximize accuracy and effectiveness.

When VGG-19 replaced ResNet-51, there was a reduction of 3.4% in accuracy, 4.3% in precision, 6.0% in recall, and 4.0% in F1 score. This indicates that ResNet-51 performs better in image feature extraction, providing richer visual features for travel review sentiment analysis.

Removing the attention mechanism significantly decreased performance across all metrics. This highlights the crucial role of attention mechanisms in the proposed method. The attention mechanism enhanced inter-modal feature correlations by leveraging text features to weight associated image features, dynamically enhancing important features while reducing the impact of noise or irrelevant information. This not only boosts sentiment analysis accuracy but also strengthens the model's generalization capacity, guaranteeing superior performance across various complex multimodal travel review settings.

Model Convergence Analysis

Figure 6 displays the loss values for all models in the ablation experiments, showing a gradual decrease as the number of iterations rises. After 20 iterations, the loss values for each model began to stabilize and converge. However, the loss values of the BERT and ResNet-51 models were always higher than those of the other models, indicating that they perform poorly in loss function optimization and converge more slowly. It is important to highlight that the ResNet-51 model's performance is constrained because it lacks sensitivity to the variability and correlation in multimodal data, preventing it from fully exploiting the useful information available. Therefore, the loss value of the ResNet-51 model was also consistently higher than that of the BERT model. The proposed method as well as the model with the removed module converged faster, and the initial loss value was lower than that of the single-modal feature extraction model, indicating that the proposed method is advantageous in dealing with multimodal data.

Figure 6. Ablation experimental model training loss values

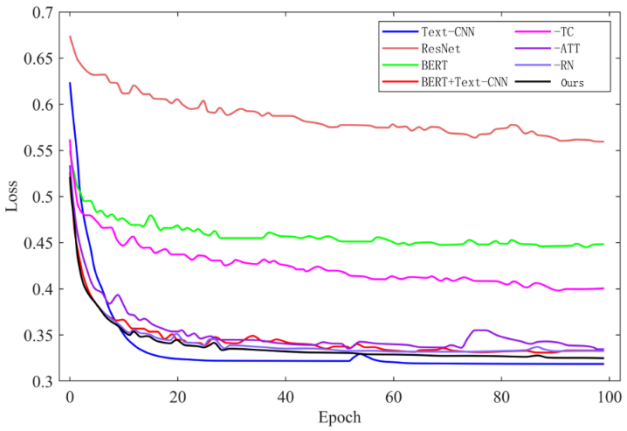


Figure 7. Accuracy of the ablation experiment model on the validation set

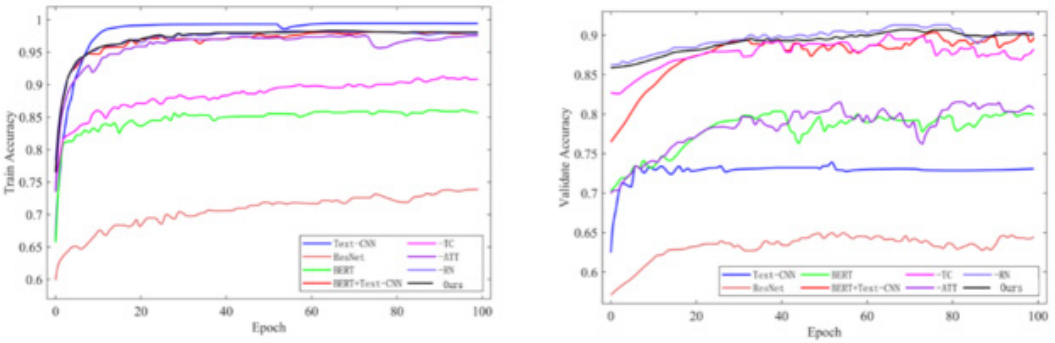


Figure 7 depicts the performance of all models across training and validation sets, comparing their outcomes. After convergence, the approach consistently outperformed other variants of the model on both sets, indicating its effective utilization of multimodal feature information and ensuring generalization performance without overfitting or underfitting. Additionally, while the Text-CNN model performed well on loss and the training set, its accuracy on the validation set was low, exhibiting typical overfitting. This indicates that the features the Text-CNN model learns from the training set do not adapt well to the validation set, which restricts its practical effectiveness. Therefore, thorough testing and validation are necessary in the model selection and optimization process to ensure generalization performance and practicality. The proposed approach better considers the differences and correlations in multimodal data, fully exploiting useful information in the data to enhance model accuracy and generalization capability.

DISCUSSION

This research presents a new method for sentiment analysis of online travel reviews, utilizing advanced models like BERT, Text-CNN, ResNet-51, and an attention mechanism. Experimental results validate its effectiveness.

Limitations

While this study makes contributions, recognizing its limitations is crucial. One of the primary limitations is the scope of the dataset used for experimentation. Although we utilized the Multi-ZOL dataset and the Ctrip Tourism Review Sentiment Analysis Dataset, these datasets may not fully represent the diversity of online travel reviews. The narrow scope of the datasets may limit how broadly the findings can be applied to various online travel platforms and user demographics.

Another limitation relates to the potential biases introduced during the data collection process. The data collected from online travel websites using data crawling may be influenced by factors such as website design, user demographics, and review filtering mechanisms. These biases could affect the representativeness of the dataset and, consequently, the interpretation of the results.

Additionally, while the algorithm demonstrates high accuracy in sentiment analysis tasks, the interpretability of the model's decisions remains a challenge. Understanding how the algorithm processes and integrates textual and visual information to arrive at sentiment predictions is crucial for ensuring the trustworthiness and usability of the system. Further research is needed to develop techniques for explaining the model's decision making process in a transparent and interpretable manner.

User Experience Analysis

In assessing the impact of the sentiment analysis algorithm, it is essential to consider its effect on the end-user experience. User experience encompasses various aspects, including usability, satisfaction, and decision making processes. By analyzing online travel reviews, the algorithm aims to provide valuable insights to end-users, such as travelers seeking information and recommendations.

First, from a usability standpoint, the effectiveness of the algorithm directly influences how users interact with online travel platforms. Accurate sentiment analysis enables users to quickly discern the overall sentiment of reviews, thereby assisting them in making informed decisions regarding travel destinations, accommodations, and attractions. Second, the impact of the algorithm on user satisfaction is crucial. By accurately capturing the sentiments expressed in travel reviews, this study aims to enhance user satisfaction by providing more relevant and personalized recommendations.

Considering the decision making processes of end-users, the sentiment analysis algorithm can significantly influence their choices. By gaining a deeper understanding of the emotional tone of reviews, users can make wiser decisions about their travel plans, leading to more enriching experiences.

Avenues for Future Research

Future studies could aim to overcome the current limitations by expanding the dataset to incorporate more diverse sources and enhancing the interpretability of algorithmic decisions. Additionally, exploring the application of the algorithm in domains beyond online travel reviews could provide valuable insights into its scalability and effectiveness.

CONCLUSION

This research presents a novel algorithm for sentiment analysis of multimodal online travel reviews, employing sophisticated models and techniques to deliver enhanced performance. The proposed approach integrates BERT and Text-CNN for sophisticated text feature extraction, ResNet-51 for detailed image feature extraction, and an attention mechanism to enhance the interplay between diverse modal data. Our method demonstrates significant improvements in accuracy and F1 scores on both the Multi-ZOL and Ctrip datasets, achieving 90.7% accuracy and a 90.8% F1 score on the Multi-ZOL dataset, and 83.6% accuracy and a 84.1% F1 score on the Ctrip dataset. These findings confirm the efficacy of merging text and image features to offer a thorough insight into tourist evaluations.

The key contributions of this research include the enhanced text feature extraction process that combines BERT's robust semantic representation with Text-CNN's proficiency in identifying essential

local attributes, and the advanced image feature extraction using ResNet-51, which significantly improves the model's predictive precision. Additionally, dynamic feature correlation was realized by using an attention mechanism to boost interactions between text and image modalities, enhancing fusion effectiveness and overall model performance.

Future research should focus on expanding the dataset to include more diverse sources and improving the interpretability of the model's decisions. Additionally, exploring the application of the algorithm in domains beyond online travel reviews could provide valuable insights into its scalability and effectiveness. Overall, this study underscores the importance of embracing multimodal approaches in sentiment analysis and demonstrates the potential of advanced algorithms to extract valuable insights from diverse sources of user-generated content.

AUTHOR NOTE

The authors declare that there are no conflicts of interest.

The data used to support the findings of this study are included within the article.

This research was supported by the Planning Project for Philosophy and Social Sciences of Hainan Province [No. HNSK(ZC)23-152].

Correspondence concerning this article should be addressed to Peng Chen, School of Tourism and Health Industry, Sanya Institute of Technology, Sanya, Hainan 572022, China. Email: cp9815202@126.com.

PROCESS DATES

Received: March 11, 2024, Revision: May 26, 2024, Accepted: May 28, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Peng Chen; cp9815202@126.com

REFERENCES

Ajitha, P., Sivasangari, A., Immanuel Rajkumar, R., & Poonguzhali, S. (2021). Design of text sentiment analysis tool using feature extraction based on fusing machine learning algorithms. *Journal of Intelligent & Fuzzy Systems*, *40*(4), 6375–6383. 10.3233/JIFS-189478

Al-Tameemi, I. K. S., Feizi-Derakhshi, M.-R., Pashazadeh, S., & Asadpour, M. (2022). A comprehensive review of visual-textual sentiment analysis from social media networks. Advance online publication. ArXiv. arXiv:2207.02160

An, S., Li, Y., Lin, Z., Liu, Q., Chen, B., Fu, Q., Chen, W., Zheng, N., & Lou, J.-G. (2022). Input-tuning: Adapting unfamiliar inputs to frozen pretrained models. Advance online publication. ArXiv. arXiv:2203.03131

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015, December 7–13). *VQA: Visual question answering*. *2015 IEEE International Conference on Computer Vision*, Santiago, Chile.

Beniwal, R., & Dobhal, A. (2024). Sentiment classification on suicide notes using Bi-LSTM model. In Swaroop, A., Polkowski, Z., Duarte Correia, S., & Virdee, B. (Eds.), *Proceedings of data analytics and management* (pp. 1–10). Springer. 10.1007/978-981-99-6547-2_1

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. (pp. 223–232.) Association for Computing Machinery. 10.1145/2502081.2502282

Chatterjee, M., Kumar, P., & Sarkar, D. (2023). Generating a Mental health curve for monitoring depression in real time by incorporating multimodal feature analysis through social media interactions. [IJIIT]. *International Journal of Intelligent Information Technologies*, *19*(1), 1–25. 10.4018/IJIIT.324600

Chen, W., Xu, Z., Zheng, X., Yu, Q., & Luo, Y. (2020). Research on sentiment classification of online travel review text. *Applied Sciences (Basel, Switzerland)*, *10*(15), 5275. 10.3390/app10155275

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Advance online publication. ArXiv. arXiv:1810.04805

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*(3), 424–444. 10.1016/j.inffus.2022.09.025

Girshick, R. (2015, December 7–13). *Fast R-CNN*. *2015 IEEE International Conference on Computer Vision*, Santiago, Chile.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Juo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *MM '17: Proceedings of the 25th ACM international conference on Multimedia* (pp. 795–816). Association for Computing Machinery. 10.1145/3123266.3123454

Kaur, R., & Kautish, S. (2022). Multimodal sentiment analysis: A survey and comparison. In I. Management Association (Ed.), *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines* (pp. 1846–1870). IGI Global. 10.4018/978-1-6684-6303-1.ch098

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In Liu, L., & White, R. (Eds.), *WWW '19: The world wide web conference* (pp. 2915–2921). Association for Computing Machinery. 10.1145/3308558.3313552

Krishnan, H., Elayidom, M. S., & Santhanakrishnan, T. (2022). A comprehensive survey on sentiment analysis in Twitter data. [IJDST]. *International Journal of Distributed Systems and Technologies*, *13*(5), 1–22. 10.4018/IJDST.300352

Li, Z., Fan, Y., Liu, W., & Wang, F. (2018). Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multimedia Tools and Applications*, *77*(1), 1115–1132. 10.1007/s11042-016-4310-5

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., & Tang, J. (2022). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Muresan, S., Nakov, P., & Villavicencio, A. (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: Short Papers, pp. 61–68). Association for Computational Linguistics. 10.18653/v1/2022.acl-short.8

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023, August). GPT understands, too. *AI Open*. 10.1016/j.aiopen.2023.08.012

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. Advance online publication. ArXiv. arXiv:1806.00064

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *MM '10: Proceedings of the 18th ACM international conference on Multimedia*. (pp. 83–92). Association for Computing Machinery. 10.1145/1873951.1873965

Momani, A. M., Alsakhnini, M., & Hanaysha, J. R. (2022). Emerging technologies and their impact on the future of the tourism and hospitality industry. [IJISSS]. *International Journal of Information Systems in the Service Sector*, *14*(1), 1–18. 10.4018/IJISSS.287579

Puh, K., & Bagić, B. M. (2023). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights*, *6*(3), 1188–1204. 10.1108/JHTI-02-2022-0078

Rao, T., Li, X., & Xu, M. (2020). Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, *51*(3), 2043–2061. 10.1007/s11063-019-10033-9

Ren, M. (2022). Research and application of the internet of things service platform based on semantic network. [IJDST]. *International Journal of Distributed Systems and Technologies*, *13*(6), 1–7. 10.4018/IJDST.308004

She, D., Yang, J., Cheng, M.-M., Lai, Y.-K., Rosin, P. L., & Wang, L. (2019). WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia*, *22*(5), 1358–1371. 10.1109/TMM.2019.2939744

Shi, P., Hu, M., Shi, X., & Ren, F. (2024). Deep modular co-attention shifting network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing Communications and Applications*, *20*(4), 1–23. 10.1145/3634706

Shi, Q., Fan, J., Wang, Z., & Zhang, Z. (2022). Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain. *Pattern Recognition*, *130*(2), 108837. 10.1016/j.patcog.2022.108837

Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(10), 13915–13916. 10.1609/aaai.v34i10.7230

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of 57th Annual Meeting of the Association for Computational Linguistics. Meeting* (pp. 6558–6569). Association for Computational Linguistics. 10.18653/v1/P19-1656

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., & Salalkhutdinov, R. (2018). Learning factorized multimodal representations. Advance online publication. arXiv. arXiv:1806.06176

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *NIPS '17: Proceedings of the 31st international conference on neural information processing* systems (pp. 6000–6010).

Veyseh, A. P. B., Lai, V., Dernoncourt, F., & Nguyen, T. H. (2021). Unleash GPT-2 power for event detection. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1*: Long Papers, pp. 6271–6282). Association for Computational Linguistics.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015, June 7–12). Show and tell: A neural image caption generator. *2015 IEEE conference on computer vision and pattern recognition*. (pp. 3156–3164). IEEE. 10.1109/CVPR.2015.7298935

Wang, G., & Shin, S. Y. (2019). An improved text classification method for sentiment classification. *Journal of Information and Communication Convergence Engineering*, *17*(1), 41–48.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 849-857). Association for Computing Machinery. 10.1145/3219819.3219903

Yang, J., Li, J., Wang, X., Ding, Y., & Gao, X. (2021). Stimuli-aware visual emotion analysis. In Macq, B. (Ed.), *IEEE transactions on image processing* (*Vol. 30*, pp. 7432–7445). IEEE.

Ye, D. D., Muthu, B. A., & Kumar, P. M. (2022). Identifying buying patterns from consumer purchase history using big data and cloud computing. [IJDST]. *International Journal of Distributed Systems and Technologies*, *13*(7), 1–19. 10.4018/IJDST.307957

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. Advance online publication. ArXiv. arXiv:1702.01923

You, Q., Jin, H., & Luo, J. (2017). Visual sentiment analysis by attending on local image regions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), 231–237. 10.1609/aaai.v31i1.10501

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 10790–10797. 10.1609/aaai.v35i12.17289

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. Advance online publication. ArXiv. arXiv:1707.07250

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Advance online publication. 10.1609/aaai.v32i1.12021

Zaw, M., & Tandayya, P. (2022). Aspect-based and multi-level sentiment information applying contrast dictionary. [IJISSS]. *International Journal of Information Systems in the Service Sector*, *14*(1), 1–22. 10.4018/IJISSS.2022010103

Zhang, C., Li, Q., & Cheng, X. (2020). Text sentiment classification based on feature fusion. *Revue d'Intelligence Artificielle*, *34*(4), 515–520. 10.18280/ria.340418

Zhang, J., Chen, M., Sun, H., Li, D., & Wang, Z. (2020). Object semantics sentiment correlation analysis enhanced image sentiment classification. *Knowledge-Based Systems*, *191*, 105245. 10.1016/j.knosys.2019.105245

Zhang, J., Liu, X., Chen, M., Ye, Q., & Wang, Z. (2022). Image sentiment classification via multi-level sentiment region correlation analysis. *Neurocomputing*, *469*, 221–233. 10.1016/j.neucom.2021.10.062

Zhang, S., Ly, L., Mach, N., & Amaya, C. (2022). Topic modeling and sentiment analysis of yelp restaurant reviews. [IJISSS]. *International Journal of Information Systems in the Service Sector*, *14*(1), 1–16. 10.4018/IJISSS.295872

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., & Sun, X. (2014a). Exploring principles-of-art features for image emotion recognition. In *MM '14 Proceedings of the 22nd ACM international conference on Multimedia* (pp. 47–56). Association for Computing Machinery. 10.1145/2647868.2654930

Zhao, S., Yao, H., Yang, Y., & Zhang, Y. (2014b). Affective image retrieval via multi-graph learning. In *MM '14: Proceedings of the 22nd ACM international conference on Multimedia* (pp. 1025–1028). Association for Computing Machinery. 10.1145/2647868.2655035

Zhao, X., Chen, Y., Li, W., Gao, L., & Tang, B. (2022). MAG+: An extended multimodal adaptation gate for multimodal sentiment analysis. In *ICASSP 2022: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4753–4757). IEEE. 10.1109/ICASSP43922.2022.9746536

Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: similarity-aware multi-modal fake news detection (2020). Advance online publication. arXiv. 200304981 10.1007/978-3-030-47436-2_27

Peng Chen, has got Master of Tourism Management from Hainan Tropical Ocean University in2020. She is currently a lecturer at Sanya Institute of Technology. Her research interests include tourism management, education management and capital market.

Lingmei Fu, has got Bachelor of Engineering from Sichuan Agricultural University in 2007. She is an associate professor at Hainan Provincial Sports Academy. Her research interests include computer applications and artificial intelligence.