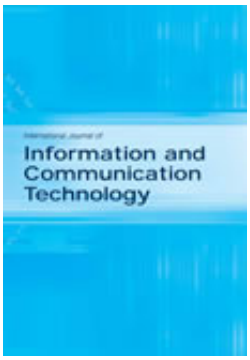
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Sentiment analysis for tourism reviews based on dual-stream graph attention fusion network

Ge Yang, Dan Han

Article History:

Received:	02 December 2024
Last revised:	20 January 2025
Accepted:	20 January 2025
Published online:	31 March 2025

Sentiment analysis for tourism reviews based on dual-stream graph attention fusion network

Ge Yang*

School of Tourism and Health Industry,
Sanya Institute of Technology,
Hainan, Sanya, 572000, China
Email: 18689979366@163.com
*Corresponding author

Dan Han

School of Business and Economics,
Sanya Institute of Technology,
Hainan, Sanya, 572000, China
Email: 13006052135@163.com

Abstract: Emotional evaluations can serve as indicators to guide the improvement of online travel service quality. This paper proposes a model based on a dual-stream graph attention fusion network. To prevent the introduction of noise unrelated to the subject matter, the network encodes the aspect texts and integrates them into the image-text embeddings through cross-attention mechanisms. Furthermore, the network incorporates a dual-stream interactive masking graph neural network to effectively align and fuse the deep features of multimodal image and text data. This allows the image data and textual data to provide rich supplementary information for each other, thereby enhancing the model's accuracy in sentiment analysis. In simulation comparisons with various state-of-the-art methods on real-world multimodal travel datasets, the results demonstrate that this approach has reduced the MAE and RMSE metrics by 19.86% and 5.26%, respectively, compared to the best-performing baseline model.

Keywords: sentiment analysis; cross-attention; interactive graph mask attention network; multimodal.

Reference to this paper should be made as follows: Yang, G. and Han, D. (2025) 'Sentiment analysis for tourism reviews based on dual-stream graph attention fusion network', *Int. J. Information and Communication Technology*, Vol. 26, No. 6, pp.117–134.

Biographical notes: Ge Yang is a Lecturer holding a Master of Science degree. He graduated from Edinburgh Napier University in the UK in 2012 and is currently employed at Sanya Institute of Technology. His academic pursuits are centred around tourism and hospitality management, as well as teaching English as a second language (TESL).

Dan Han is an Associate Professor with a Master's in Computer Science and Technology. She completed her studies at Wuhan University in 2010 and is currently affiliated with Sanya Institute of Technology. Her research focuses on data analysis and software development.

1 Introduction

The rapid evolution of internet technology has made social media and online platforms pivotal conduits for individuals to share travel experiences, articulate emotions, and exchange information (Huangxiong et al., 2022). The travel reviews generated by users not only offer invaluable insights for prospective travellers but also present the tourism industry with opportunities to discern consumer preferences and demands. The wealth of emotional information embedded in these reviews is crucial for enhancing travel operations and delivering personalised services (Ghazali et al., 2021; Zheng et al., 2023).

The initial sentiment analysis in the domain of travel reviews was primarily conducted on brief textual data. The predominant analytical frameworks encompassed techniques like SVM-based analysis (Chen, 2014), RNN-driven approaches (Xiang et al., 2023; Guo and Jia, 2020), CNN-structured methods (Du, 2022), and BERT-leveraged strategies for sequential data examination (Zhu et al., 2024).

Monomodal sentiment analysis has limitations in terms of information volume, contextual supplementation, dynamic capture, causality recognition, and simulation of interactions, whereas multimodal sentiment analysis can provide more comprehensive and accurate results by integrating information from different modalities (Zhang et al., 2023a). As a result, multimodal sentiment analysis algorithms for tourism reviews have become a research hotspot. However, multimodal sentiment analysis algorithms face two key challenges:

1 multimodal data may contain more noise information, and eliminating this noise to prevent negative impacts on sentiment analysis results is a challenge

2 effectively fusing the features of different modalities in multimodal sentiment analysis, preserving the semantic integrity of each modality, and achieving robust integration between them is a significant challenge.

To address these issues, this paper introduces a dual-stream graph attention fusion network for tourism review sentiment analysis, with the following innovations:

1 In response to the scarcity of sentiment analysis corpora in the tourism domain that integrates images and text, this paper constructs a dataset by scientifically scoring the sentiment analysis of six aspects of travel image-text reviews collected from the web, combined with the project background.

2 During the feature extraction process, the multi-head cross-attention mechanism within the dual-stream framework fuses aspect-class embeddings with image-text embeddings, effectively avoiding the introduction of noise from image-text information unrelated to the aspect.

3 During the feature extraction process, a dual-stream multi-layer graph attention network is utilised to integrate deep features of image-text pairs. In this approach, each stream branch employs an adjacency matrix derived from the dot product of image-text features; however, the order of the dot product differs.

2 Literature review

In the realm of aspect sentiment analysis, the research areas closely related to it primarily include text sentiment analysis and multimodal sentiment analysis. The following sections will introduce the current state of research in these three domains over recent years.

2.1 Text sentiment analysis

The task of sentiment analysis in text involves inferring the emotional polarity associated with a specific aspect category from a given text. In this process, recurrent neural networks and conventional attention mechanisms are predominantly employed to autonomously acquire semantic characteristics of both the context and the aspectual terms. Tang et al. (2015) proposed a bidirectional target-dependent sentiment classification model based on LSTM, which is capable of considering contextual information from both left and right directions, thereby better comprehending the semantic relationships and emotional tendencies within a sentence. Tang et al. (2016) proposed a deep memory network (MemNet) for aspect-level sentiment classification that deduces emotional tendencies by calculating the significance of each contextual word. Compared to SVM and LSTM models, this approach shows a significant enhancement in efficiency. Ma et al. (2017) introduced an interactive attention network (IAN) for aspect-level sentiment classification that leverages interactive learning to concurrently focus on context and target, generating distinct representations for both, which enhances the model's ability to grasp the relationship between target words and context. Chen et al. (2017) proposed a recurrent attention network on memory (RANM) for aspect-level sentiment analysis. This model enhances its ability to capture emotional features within complex sentence structures by nonlinearly integrating the outputs of multiple attention layers. Gu et al. (2018) utilised an attention mechanism to focus on the semantic relationship between aspect terms and sentences. Song et al. (2019) proposed a novel model that can circumvent the use of RNNs and employs an attention mechanism to model the relationship between context and targets. However, these approaches have not taken into account the interaction between aspects and context. Gao et al. (2019) introduced three BERT-based models that leverage BERT's powerful semantic representation capabilities and the positional information of target words to obtain target-related feature vectors. By adding the target word as a special token to the input sequence, the model enhances the target information, thereby improving the effectiveness of sentiment classification. Xu et al. (2020) employed global and local attention to capture different granularity interaction information between aspects and context. Wu et al. (2020) proposed a relative position encoding layer that integrates the positional information of a given aspect word and uses an aspect attention mechanism to consider the semantic relationship between aspect words and context words. Alshuwaier et al. (2022) focused on a systematic literature review of deep learning methods for emotion analysis based on documents, providing a brief overview of the latest developments in emotion analysis technology and the application of deep learning in the field of document analysis, as well as assessments and enhancements. Shi et al. (2023) proposed a joint learning approach for cross-domain aspect term extraction based on soft prompts, which combines external linguistic features and employs multi-objective learning to bridge the gap between domain-invariant representations of source and target domains,

compensating for the disparities in aspect term distributions across different domains. Atchariyachanvanich et al. (2024) employed an enhanced classification model that integrates XGBoost and Lasso techniques, augmenting it with 33 innovative features – including textual elements, sentiment-informed characteristics, and dictionary-based attributes such as herbal medicines, fruits, and vegetables – to pinpoint the objectives of articles. This strategy aims to elevate the precision of detecting counterfeit news in the health and medical sectors on websites that are in the Thai language. Luo (2024) introduced a text-sensitive analysis model known as BERT-CNN-AbiLSTM. This model refines the TextCNN component, enhancing the extraction of local features and optimising the feasibility of the model. Li (2024) proposed an emotion analysis method based on a teaching evaluation sentiment dictionary. This approach incorporates parallel relationships and similarity calculations, and designs a multi-level polarity recognition strategy for emotional words, thereby enhancing the performance of sentiment analysis.

In recent years, research on tourism has proliferated both domestically and internationally, and sentiment analysis based on text within the tourism field has garnered the attention of scholars. For instance, Paolanti et al. (2021) introduced a joint multi-grain topic sentiment model that effectively captures consumers' emotional tendencies by analysing multiple semantic levels of online reviews simultaneously. Borrajo-Millán et al. (2021) introduced a method leveraging text mining and sentiment analysis to evaluate the performance of tourism destinations and explores its implications for the sustainability of these destinations. Fang et al. (2022) presented an ELECTRA-based model for sentiment analysis of tourist attraction reviews, enhancing the accuracy of sentiment classification by leveraging ELECTRA's ability to capture contextual and target-specific semantic features.

2.2 Multimodal sentiment analysis

Multimodal sentiment analysis through the integration of various modalities can provide more comprehensive and accurate results than single-text sentiment analysis, becoming the mainstream of today's sentiment analysis algorithms. In the realm of image-text sentiment analysis, the integration strategy of visual and textual elements is crucial for addressing the challenges of sentiment analysis tasks. Numerous early studies (Poria et al., 2016; You et al., 2017) employed feature-level fusion for learning image-text features, which simply concatenated the two types of features to form a longer feature vector. However, this approach fails to capture the interrelations and differences between images and text, potentially losing some valuable information. To address this issue, researchers have initiated the use of decision fusion and hybrid integration methods to combine multimodal features, for instance (Truong and Lauw, 2019) that visual information aids in pinpointing significant sentences within textual content. Consequently, they leverage an attention mechanism to employ image data as an alignment tool to highlight crucial sentences in the document, introducing an attention network focused on the visual aspect. Chen et al. (2018) introduced a deep fusion convolutional neural network designed to concurrently learn emotional representations from both text and images, merging the two modalities at the pooling layer. Xu et al. (2019a) proposed a bidirectional multi-level attention model (BDMLA) that integrates regional features of images with multiple semantic levels of text, utilising an attention mechanism to align the important sentences in the document. This model employs both an image attention network and a semantic attention network, thereby effectively aligning

the multimodal information. Xu et al. (2019b) proposed a multimodal sentiment analysis task that focuses on image and text, and introduces a multi-interactive memory network model (MIMN) that captures interactions between the two modalities and within a single modality, but overlooks the characteristics of image-text data. Nan et al. (2018) introduce a multimodal sentiment analysis approach that utilises a co-memory network to integrate textual and visual information, thereby enhancing the precision of sentiment recognition. Yu and Jiang (2019) enhanced BERT's capabilities for target-oriented sentiment analysis by adding a target attention mechanism to align textual targets with visual content from images, resulting in improved sentiment classification for targets in multimodal data. Khan and Fu (2021) proposed a method that leverages object-aware transformers to translate images and generate text, which is then used to enhance the sentiment analysis of targets in multimodal data, achieving state-of-the-art performance without modifying BERT's architecture for multimodal inputs. Wang et al. (2023) introduced an efficient multimodal transformer that enhances the accuracy of multimodal sentiment analysis through adaptive modality weighting. This study incorporates a multimodal adaptive weighting matrix to allocate appropriate weights to each modality based on its contribution to sentiment analysis. Zhang et al. (2023a) introduced an innovative approach to multimodal sentiment analysis known as the adaptive modality-specific weight fusion network (AdaMoW), aimed at enhancing the precision of multimodal data integration. During the training phase, this method leverages correlation analysis to assign weights to each modality, employing a weight-mapping network to learn these weights. Li et al. (2024) introduced an innovative approach to multimodal sentiment analysis known as the adaptive token selection and fusion network (ATSFN), which aims to enhance the accuracy of sentiment analysis by dynamically selecting and fusing informative tokens from multimodal data.

In recent years, multimodal sentiment analysis has also been widely applied in the field of tourism reviews, for instance (Zhang et al., 2023b) introduced a multimodal sentiment analysis method (TBGAV) based on images and text, which enhances the accuracy of sentiment analysis in tourism reviews through three modules: image sentiment extraction, text sentiment extraction, and image-text fusion. Chen and Fu (2024) introduced an advanced multimodal sentiment analysis approach for tourism reviews, which combines BERT and Text-CNN models for text feature extraction and utilises ResNet-51 for image feature extraction, while incorporating an attention mechanism to enhance the correlation between modalities.

This paper also takes the multimodal data of text and images as the basis, designing a dual-stream graph attention fusion network for sentiment analysis of travel reviews. This method not only effectively utilises aspect-based embedding information to eliminate noise in text and images but also introduces a novel multimodal fusion approach: the dual-stream interactive masked graph attention network. This approach allows image data and text data to modulate each other's hidden features through a masked adjacency matrix, thereby unearthing more useful complementary information.

3 Dataset construction

In the current landscape, research into sentiment analysis within the tourism sector, especially concerning visual and textual content, is only just beginning. The dearth of corpus resources is a pivotal issue confronting such studies. To implement sentiment

analysis focused on visual and textual elements in the tourism industry, it is imperative to gather and annotate pertinent data, thereby improving the model's capacity to understand and categorise the emotions expressed in reviews. This article offers a valuable data asset for research in this area by establishing a dataset for sentiment analysis that encompasses visual and textual aspects of tourism. The specific steps are as follows:

1 Dataset crawling and preprocessing.

   By scraping online travel websites, a dataset of travel-related image-text reviews was constructed, comprising 1,467 tourist attractions and 42,450 image-text reviews. Subsequently, duplicate reviews and repetitive descriptions within the reviews were eliminated, retaining only unique entries. Additionally, comments unrelated to travel and irrelevant symbols in the data were removed.

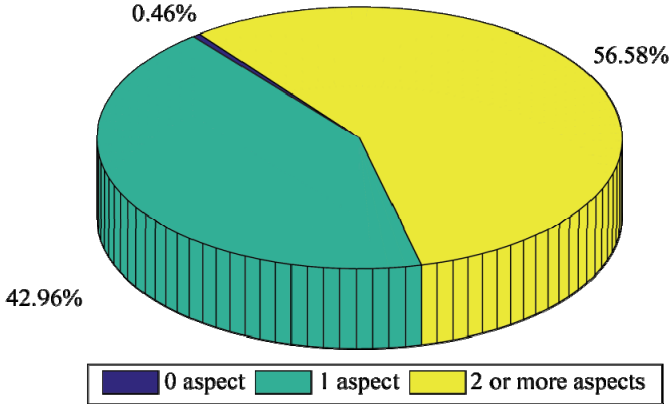
2 Determine the category of aspects

   Through the analysis of travel reviews, it was found that some comments not only contain content from multiple aspects but may also exhibit differences in emotional tendencies. Therefore, by analysing the data from travel reviews, this paper defines six aspect categories: ambiance, admission fee, management, transportation, entertainment, and recommendability.

3 Data annotation

   For the data preprocessed in Step 1, manual annotation was conducted. The aspects involved were tagged and statistically analysed. The results are presented in Figure 1.

Figure 1 Tourism review aspect statistics (see online version for colours)



As can be seen from Figure 1, sentences containing two or more aspects account for 56.58% of the total dataset, those with a single aspect represent 42.96% of the total dataset, and those with missing aspects constitute 0.46% of the total dataset. This paper employs a manual annotation method to label the aspect categories and sentiment polarity in travel reviews. If a review contains n aspect categories, the review is duplicated n times, with only one aspect category retained in each copy and its corresponding sentiment polarity annotated. To more finely express the emotional intensity of the reviews, we rate the emotions for each aspect on a scale from 0 to 1, with the strength of

emotional sentiment represented by different scores. The specific implementation involves five annotators who independently label the data. After the annotation is completed, the maximum and minimum values are discarded, and the average of the remaining three values is taken as the emotional score for each aspect.

4 Methodology

The task of aspect-based sentiment analysis within the multimodal context of text and images can be formally defined as: Let the textual input features be represented by $T = (T_1, T_2, \cdots, T_l)$ and the visual input features by $I = (I_1, I_2, \cdots, I_p)$. The goal is to develop a model that forecasts the sentiment score for a specified aspect term $A = (A_1, A_2, \cdots, A_c)$. In this context, l indicates the number of text input features, p denotes the number of image input features, and c is the number of aspects. The overall architecture of the dual-stream graph attention fusion network's aspect-based sentiment analysis model designed in this paper is depicted in Figure 2, encompassing the multimodal feature acquisition, multimodal feature fusion and sentiment analysis classifier.

4.1 Multimodal feature acquisition

For extracting text features, Yake, an efficient keyword extraction tool, first identifies the top l keywords. It assigns them a score that takes into account n-gram frequency, the terms' distribution rarity, and their positional weight. Following this process, the initial text embedding is obtained by embedding each keyword using the pre-trained BERT model:

$$H_{T_l} = BERT(T_l) \tag{1}$$

where T_l represents the original input feature for the l^{th} keyword, while H_{T_l} denotes the corresponding embedding for the same keyword in the text. BERT, which stands for bidirectional encoder representations from transformers. Central to the BERT model is the transformer architecture, which learns deep text representations through extensive pre-training on large text corpora (Devlin et al., 2019).

For extracting image features, the initial step involves dividing the image into several fixed-sized segments through a patching technique, facilitating subsequent operations like feature extraction or analysis. Given an original image dimensioned at (H, W, C), with each segment, or patch, measuring (h, w, c) and a stride of s, the equation to determine the total number of resultant patches, denoted as p, is as follows:

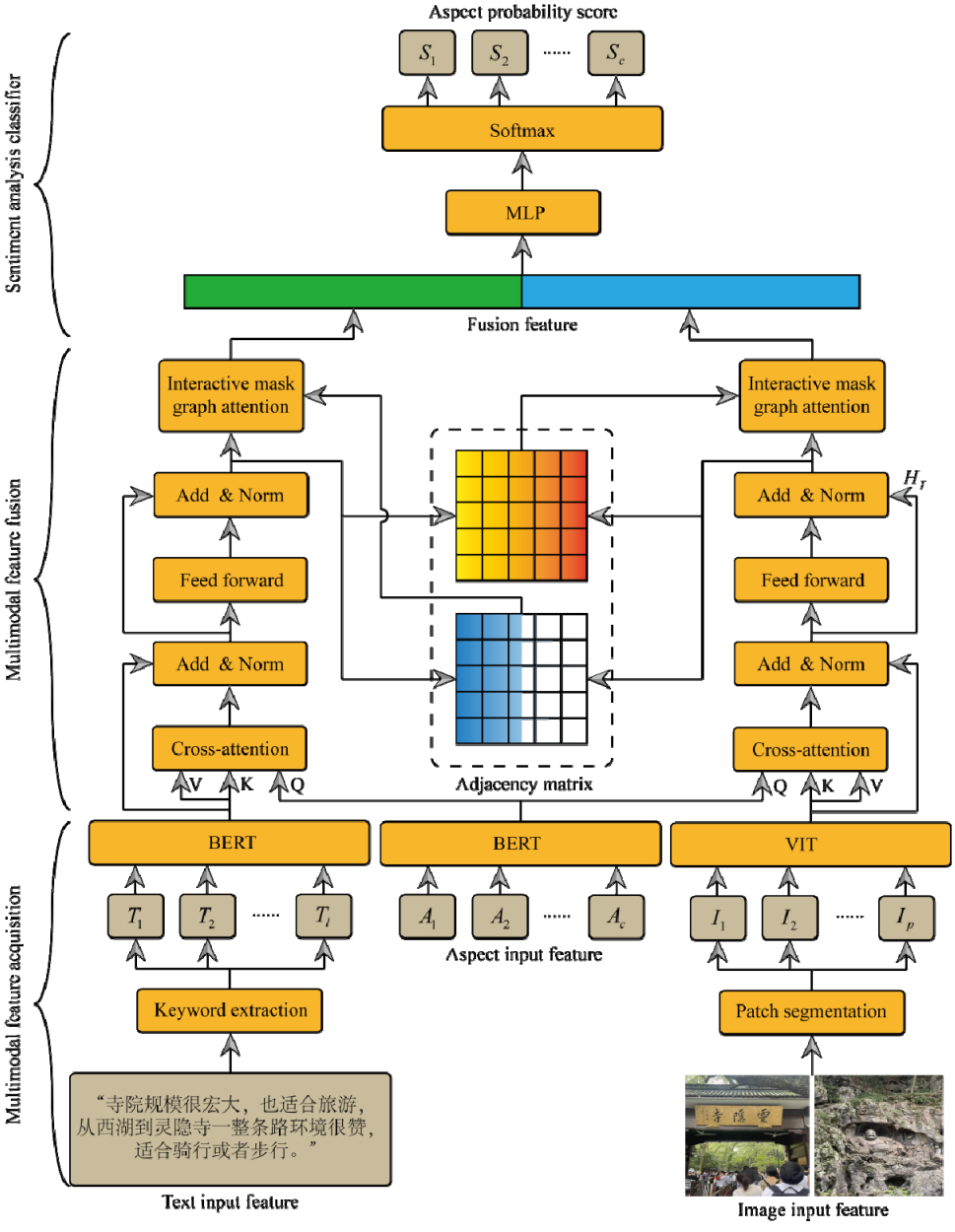
$$p = \frac{(H-h)(W-w)}{(s+1)^2} \tag{2}$$

To acquire the initial image embedding, each patch's features are fed into the trained vision transformer (VIT) model (Dosovitskiy et al., 2020):

$$H_{I_p} = VIT(I_p) \tag{3}$$

where I_p denote the original feature of the image's p^{th} patch, while H_{I_p} represents its corresponding embedding.

Figure 2 The structure of dual-stream graph attention fusion network (see online version for colours)



For extracting aspect features, this paper still carries out initial text embedding for each aspect through BERT network:

$$H_{A_c} = BERT\left(A_c\right) \tag{4}$$

where A_c represents the original input feature for the c^{th} aspect, while H_{A_c} denotes the corresponding embedding for the same aspect.

4.2 Multimodal feature fusion

After obtaining the initial embeddings of multimodal features, this paper first integrates the aspect category embedding with text embedding and image embedding respectively through a dual-stream cross-attention network.

$$\begin{cases} Z_T = Soft\max\left(\dfrac{H_A W_T^Q \left(H_T W_T^K\right)^T}{\sqrt{D_T^K}}\right) H_T W_T^V \\ Z_I = Soft\max\left(\dfrac{H_A W_I^Q \left(H_I W_I^K\right)^T}{\sqrt{D_I^K}}\right) H_I W_I^V \end{cases} \tag{5}$$

where Z_T and Z_I are respectively the text features and image features that have integrated aspect embedding; $W_T^Q, W_T^K,$ and W_T^V are respectively the weights of the linear layers for cross-attention on the text branch; while $W_I^Q, W_I^K,$ and W_I^V are the weights of the linear layers for cross-attention on the image branch; $1/\sqrt{D_T^K}$ and $1/\sqrt{D_I^K}$ are the scaling factors for cross-attention on the text and image branches, respectively. By determining the correlation between aspect-based features and both textual and visual features through dot product, it is possible to effectively suppress irrelevant text-image information.

Subsequently, the integrated textual and visual features are processed through two-layer normalisation and a feedforward neural network, with each layer normalisation followed by a residual shortcut:

$$\begin{cases} E_T = Norm\left[Norm\left(Z_T + T\right)\left(W_T + 1\right)\right] \\ E_I = Norm\left[Norm\left(Z_I + I\right)\left(W_I + 1\right)\right] \end{cases} \tag{6}$$

where T and I denote initial text embedding and image embedding respectively, $Norm(\cdot)$ denote the layer normalisation operation, which is intended to mitigate internal covariate shift and expedite the rate of convergence:

$$Norm(X) = \frac{X - E(X)}{\sqrt{Var(X) + \varepsilon}}\gamma + \beta \tag{7}$$

where $E(X)$ and $Var(X)$ represent the expected value and variance of the input feature X, while γ and β are the learnable scaling and shifting parameters.

Finally, a dual-stream interactive mask graph attention network is employed to integrate deep textual and visual features from both branches. In the textual branch, the hidden features of all key terms form a graph structure, and in the image branch, the hidden features of all small patch images form a graph structure as well. The adjacency matrix is obtained through the dot product of the hidden features from the two branches, but the order of the dot product is reversed.

$$\begin{cases} G_T = Soft\max\left(\dfrac{E_T W_T^G \left(E_T W_T^G\right)^T}{\sqrt{D_{E_T}}} \otimes Mask\left(Sigmoid\left(E_T E_I^T\right)\right)\right) E_T W_T^G \\ G_I = Soft\max\left(\dfrac{E_I W_I^G \left(E_I W_I^G\right)^T}{\sqrt{D_{E_I}}} \otimes Mask\left(Sigmoid\left(E_I E_T^T\right)\right)\right) E_I W_I^G \end{cases} \tag{8}$$

where G_T and G_I denote the output features of dual-stream interactive mask graph attention network, W_T^G and W_I^G denote the linear weights of the graph attention network in the text branch and the image branch respectively, $E_T E_I^T$ and $E_I E_T^T$ respectively obtain the adjacency matrices of the text and image graph structures, $Mask(\cdot)$ refers to the masking operation:

$$Mask\left(X[i, j]\right) = \begin{cases} X[i, j], & \text{if } X[i, j] \geq \zeta \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where ζ indicates a mask threshold.

4.3 Sentiment analysis classifier

After acquiring dual-stream multimodal fusion features, this paper utilises a sentiment analysis classifier to obtain the final aspect score. Initially, the dual-stream fusion features are concatenated; subsequently, a multilayer perceptron is employed to adjust the dimension of the concatenated features, aligning it with the number of categories in the aspect. Finally, a Softmax activation function is applied to derive the scores for each aspect:

$$S = Soft\max\left[MLP\left(G_T \| G_I\right)\right] \tag{10}$$

5 Experimental analysis

All experiments in this article were conducted on a computer equipped with an Nvidia GeForce RTX3090 GPU.

5.1 Parameter setting

For images, this paper adjusts their size to 224 × 224 pixels, using 32 × 32 pixels as the size of each patch, resulting in 49 patches, and utilises VIT to obtain a 49 × 256-dimensional image feature vector; for the text model, the number of keywords is selected to be 16, and BERT is used to obtain a 16 × 256-dimensional text feature vector; for the aspect vector, BERT is employed to acquire a 6 × 256-dimensional aspect feature vector; for the dataset, it is divided in an 8:1:1 ratio to obtain the training set, validation set, and test set; for training hyperparameters, the learning rate is set to 1e-5, the batch size is set to 4, dropout is set to 0.3, the number of attention heads is set to 2, and the optimiser used is Adam.

5.2 Baseline models setting

To validate the superiority of our designed approach, we will compare it with eight advanced baseline models that are deeply studied, including LSTM (Tang et al., 2015), MemNet (Tang et al., 2016), IAN (Ma et al., 2017), and RAMN (Chen et al., 2017) as unimodal text models; co-memory (Nan et al., 2018), MIMN (Xu et al., 2019a), AdaMoW (Zhang et al., 2023b) and ATSFN (Li et al., 2024) as multimodal image-text models, with detailed descriptions of each baseline model provided in Table 1.

Table 1 The detailed descriptions of each baseline model

Text-only modality	LSTM	Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that is capable of learning long-term dependencies in sequential data. It is particularly effective for text sentiment analysis as it can capture contextual information over long sequences.
	MemNet	Memory network (MemNet) is a deep learning model that uses a memory component to store and retrieve information. It explicitly considers the importance of each contextual word in determining the sentiment polarity of a given aspect.
	IAN	Interactive attention network (IAN) models both the target (aspect) and context separately and uses attention mechanisms to interact between them. This allows the model to generate target-related context representations and context-related target representations.
	RAMN	Recurrent attention network on memory (RAMN) is a memory-based model that uses attention mechanisms to capture the relationship between the target and context. It nonlinearly integrates multiple attention layers to enhance sentiment classification.
Image-text multimodality	Co-memory	Co-memory network uses two memory units to store visual and textual features separately. It employs attention mechanisms to update and access the memory, allowing for the integration of multimodal information.
	MIMN	Multi-interactive memory network (MIMN) includes three memory units for aspect, text, and image features. It uses attention mechanisms and nonlinear combinations to update and read memory, capturing interactions between modalities.
	AdaMoW	Adaptive modality-specific weight fusion network (AdaMoW) adaptively assigns weights to each modality based on its contribution to sentiment analysis. It uses a weight-mapping network to learn the optimal weights for each modality.
	ATSFN	Adaptive token selection and fusion network (ATSFN) dynamically selects and fuses informative tokens from multimodal data. It estimates the unique contribution of each modality to emotional tendencies, enhancing the accuracy of sentiment analysis.

To validate the superiority of our proposed dual-stream graph attention fusion network, we compare it with eight advanced baseline models. These models were selected to represent a range of approaches, from traditional text-only models (LSTM, MemNet, IAN, RAMN) to state-of-the-art multimodal models (co-memory, MIMN, AdaMoW, ATSFN). The text-only models provide a baseline for understanding the improvements

brought by incorporating multimodal data, while the multimodal models demonstrate the effectiveness of different fusion strategies. By comparing our model to these baselines, we can comprehensively evaluate its performance in terms of both accuracy and robustness.

5.3 Performance superiority analysis

Commencing with an examination of the accuracy for all recommendation models, the mean absolute error (MAE) and root mean square error (RMSE) are the parameters for gauging performance.

$$\begin{cases} MAE = \dfrac{1}{E_{test}} \sum_{i=1}^{E_{test}} \left| y_i - \hat{y}_i \right| \\ RMSE = \sqrt{\dfrac{1}{E_{test}} \sum_{i=1}^{E_{test}} \left(y_i - \hat{y}_i \right)^2} \end{cases} \tag{11}$$

where E refers to the count of samples in the test set, where y is the true score and y' is the estimated score. The performance of the model is gauged by the MAE and RMSE, which quantify the prediction errors; lower values of these metrics indicate higher model accuracy

Table 2 The sentiment score error on different models

Data type	*Model*	*MAE*	*RMSE*
Text	LSTM	2.231	2.753
	MemNet	1.691	2.145
	IAN	1.617	2.053
	RAMN	1.749	2.261
Text+Image	Co-Memory	1.323	1.608
	MIMN	1.271	1.537
	AdaMoW	1.202	1.457
	ATSFN	1.183	1.312
	Ours	0.948	1.243

From Table 2, it is evident that LSTM performs the poorest, likely due to its failure to leverage aspect-category word information. In contrast, MemNet, IAN, and RAMN utilise attention mechanisms to extract information from the given aspect text. Compared to single-text modal methods, the model designed in this paper integrates image data, allowing for a multi-level fusion of aspect information and image-text information, thereby preserving more and more important aspect-category sentiment information.

The performance of the model proposed in this paper also surpasses other methods that incorporate image data. The Co-memory method and the MIMN method are similar to MemNet, but the MIMN method uses an aspect-guided attention mechanism to generate attention vectors for text and images, which performs better than co-memory. AdaMoW employs an adaptive weight fusion strategy, while ATSFN utilises an adaptive token selection strategy to adaptively select and integrate information from different modalities to enhance the accuracy of sentiment analysis. Compared to other baseline

models, their performance tends to be better. The model designed in this paper not only effectively integrates aspectual features into the visual-textual features but also reduces the redundant information in the individual deep features of images and text through a novel graph neural network, thereby achieving higher accuracy compared to these baseline models.

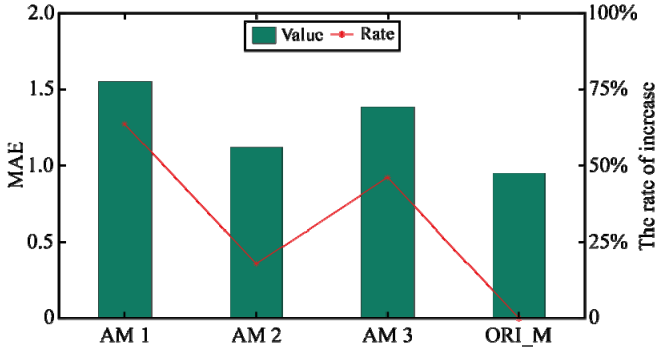
5.4 Ablation experiment

To verify the performance of various components of the model, this paper designs three ablation models for performance comparison with the original model. The specific descriptions of the ablation models are shown in Table 3.

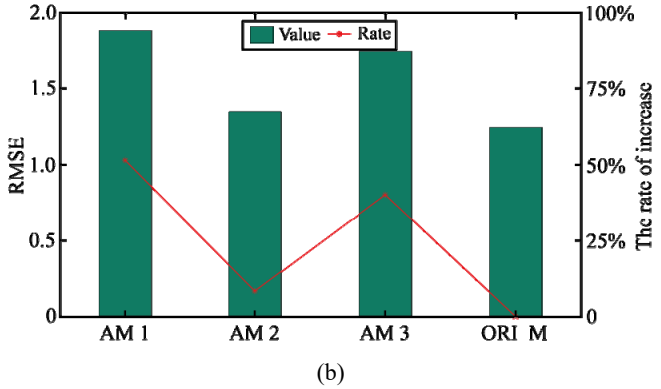
Table 3 The detailed descriptions of each ablation model

Ablation model 1	Based on the original model, the branch for image data processing is removed, relying solely on the text unimodality for feature extraction.
Ablation model 2	Based on the original model, the aspect-category interaction with image-text features is removed, and the cross-attention is replaced with self-attention.
Ablation model 3	Based on the original model, the inter-branch interaction mask attention network component is removed.

Figure 3 The score errors of the model vary with the number of patches, (a) the MAE (b) the RMSE (see online version for colours)



(a)



(b)

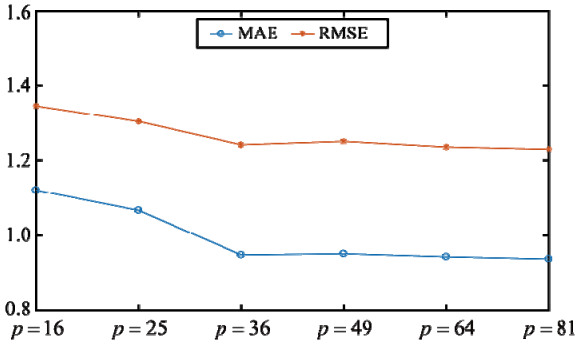
From Figure 3, it is evident that the error for each ablation model increases compared to the original model, indicating that each removed component plays a positive role in the model's performance. The increase in error is highest for ablation model 1, which removes the image data processing branch, suggesting that the visual modality provides significant supplementary information to the text, enhancing the model's ability to capture sentiment more accurately. The increase in error is lowest for ablation model 2, which removes the aspect-category interaction with image-text features, indicating that while the cross-attention mechanism is important, its impact on performance is relatively less compared to the image data processing branch. Finally, ablation model 3, which removes the inter-branch interaction mask attention network, shows a moderate increase in error, highlighting the importance of this component in dynamically aligning and integrating features from both modalities.

5.5 Key parameter selection

The innovative approach presented in this paper is characterised by three fundamental parameters: the number of patches p, the number of keywords l, and the mask threshold ζ. Figure 4's panels (a), (b), and (c) respectively depict the model's error as p varies, the model's error as l varies, and the model's error as ζ varies.

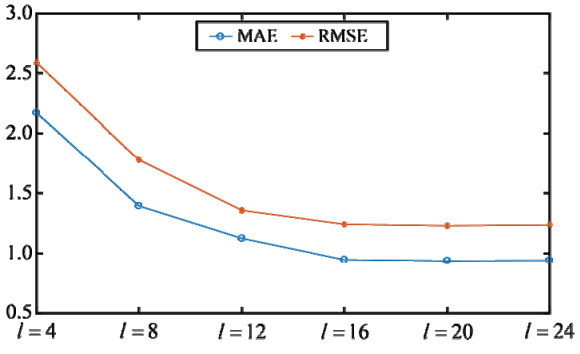
Initially, it can be observed that as p and l increase, the model's error gradually decreases. However, when p reaches 36 or l reaches 16, the rate of error reduction becomes very slow. This indicates that the model's performance plateaus within a certain range for p and l, and further increasing these values to avoid computational waste is unnecessary. Conversely, as ζ increases, the model's error also increases incrementally. The error does not change significantly when ζ rises from 0.1 to 0.3. However, when ζ continues to increase beyond 0.3, the rate of error increase accelerates. This suggests that a minor increment in e has a negligible impact on the method's analytical performance. A substantial increase in ζ, however, can lead to more information being obscured, resulting in a decline in model performance.

Figure 4 The score errors of the model with the change in different parameters, (a) with the change in the number of patches (b) with the change in the number of keywords (c) with the change in the number of mask threshold (see online version for colours)

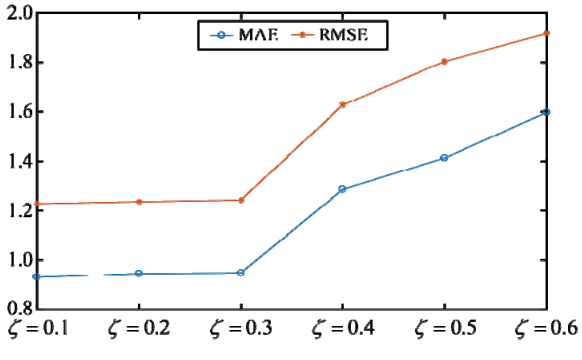


(a)

Figure 4 The score errors of the model with the change in different parameters, (a) with the change in the number of patches (b) with the change in the number of keywords (c) with the change in the number of mask threshold (continued) (see online version for colours)



(b)



(c)

6 Conclusions

This paper addresses the issue of data scarcity in the tourism domain's image-text sentiment analysis task by crawling web data to construct a tourism image-text aspect-based sentiment analysis dataset. In response to the image-text aspect-based sentiment analysis task, this chapter proposes a dual-stream graph attention fusion network that deeply explores the interrelationships and influences between aspects and image-text data. Initially, aspect information is integrated with image-text information through a cross-attention mechanism. Subsequently, an interactive mask graph attention network is utilised to deeply fuse image-text features. This approach not only enables image-text data to provide rich supplementary information to each other but also prevents the introduction of aspect-irrelevant information. Experiments conducted on real datasets demonstrate that the proposed model outperforms other comparative models.

Declaration of interest

The author declares that it does not have any known interests or personal relationships that could potentially influence the reported work in this paper.

Acknowledgements

This research was supported by Hainan Provincial Higher Education Scientific Research Project, 'Research on the Optimisation of Hainan Road Traffic Governance System Based on Big Data Technology', Project Number: hnky2024ZD-25, Sanya Federation of Social Sciences Circles Project, 'Research on the Optimisation of Road Traffic Congestion Governance Model in Sanya City from the Perspective of Big Data', Project Number: SYSK2023-03, 2023 Research and Innovation Team Project of Sanya Institute of Technology and Vocational College, 'Research on Digital Technology Empowering Smart City Construction under the Background of Hainan Free Trade Port', Project Number: SITKY2023-01.

References

Alshuwaier, F., Areshey, A. and Poon, J. (2022) 'Applications and enhancement of document-based sentiment analysis in deep learning methods: systematic literature review', *Intelligent Systems with Applications*, September, Vol. 15, p.200090.

Atchariyachanvanich, K., Saengkhunthod, C., Kerdnoonwong, P., Chanlekha, H. and Cooharojananone, N. (2024) 'Improvement of a machine learning model using a sentiment analysis algorithm to detect fake news', *Journal of Cases on Information Technology*, Vol. 26, No. 1, pp.1–26.

Borrajo-Millán, F., Alonso-Almeida, M.D., Escat-Cortés, M. and Yi, L. (2021) 'Sentiment analysis to measure quality and build sustainability in tourism destinations', *Sustainability*, Vol. 13, No. 11, p.6015.

Chen, L. (2014) 'A cost adjusting method for increasing customers' sentiment classification performance', *International Journal of Information and Electronics Engineering*, Vol. 4, No. 5, pp.336–339.

Chen, P. and Fu, L. (2024) 'Enhancing multimodal tourism review sentiment analysis through advanced feature association techniques', *International Journal of Information Systems in the Service Sector*, Vol. 15, No. 1, pp.1–21.

Chen, P., Sun, Z., Bing, L. and Yang, W. (2017) 'Recurrent attention network on memory for aspect sentiment analysis', *Conference on Empirical Methods in Natural Language Processing*.

Chen, X., Wang, Y. and Liu, Q. (2018) 'Visual and textual sentiment analysis using deep fusion convolutional neural networks', *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, Beijing, China, pp.1557–1561.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'Bert: pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp.4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020) *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ArXiv, abs/2010.11929.

Du, C. (2022) 'Research and implementation of emotional analysis and visualization system for tourist attraction comments', *Information and Computer*, Vol. 24, pp.154–157.

Fang, H., Xu, G., Long, Y. and Tang, W. (2022) 'An effective ELECTRA-based pipeline for sentiment analysis of tourist attraction reviews', *Applied Sciences*, Vol. 12, No. 21, p.10881.

Gao, Z., Feng, A., Song, X. and Wu, X. (2019) 'Target-dependent sentiment classification with BERT', *IEEE Access*, Vol. 7, pp.154290–154299, https://doi.org/10.1109/ACCESS.2019.2946594.

Ghazali, R., Rashid-Radha, J. and Mokhtar, M. (2021) 'Tourists' emotional experiences at tourism destinations: analysis of social media reviews', *Journal of Social Media Reviews*, Vol. 1, No. 1, pp.49–70.

Gu, S., Zhang, L., Hou, Y. and Song, Y. (2018) 'A position-aware bidirectional attention network for aspect-level sentiment analysis', *International Conference on Computational Linguistics*.

Guo, Q. and Jia, G. (2020) 'Research on sentiment analysis method of travel reviews based on tree LSTM', *Application Research of Computers*, Vol. 37, No. S2, pp.63–65.

Huangxiong, Q., Rucong, M., Juan, X., Changchun, J. and Huili, C. (2022) 'The relationship between social media features, perceived image and tourists' behavioral intentions a case study of Chongqing', in *West Forum on Economy and Management*, Vol. 33, No. 1, pp.1–8.

Khan, Z. and Fu, Y. (2021) 'Exploiting BERT for multimodal target sentiment classification through input space translation', *Proceedings of the ACM MM*, ChengDu, China, pp.3034–3042.

Li, C. (2024) 'Sentiment classification model for student teaching evaluation based on deep learning technology', *International Journal of Information and Communication Technology*, Vol. 24, No. 7, pp.1–16.

Li, X., Lu, M., Guo, Z. and Zhang, X. (2024) 'Adaptive token selection and fusion network for multimodal sentiment analysis', *Conference on Multimedia Modeling*.

Luo, Z. (2024) 'A study into text sentiment analysis model based on deep learning', *International Journal of Information and Communication Technology*, Vol. 24, No. 8, pp.64–75.

Ma, D., Li, S., Zhang, X. and Wang, H. (2017) *Interactive Attention Networks for Aspect-Level Sentiment Classification*, ArXiv, abs/1709.00893.

Nan, X., Mao, W. and Chen, G. (2018) 'A co-memory network for multimodal sentiment analysis', *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.929–932.

Paolanti, M., Mancini, A., Frontoni, E., Felicetti, A., Marinelli, L., Marcheggiani, E. and Pierdicca, R. (2021) 'Tourism destination management using sentiment analysis and geo-location information: a deep learning approach', *Information Technology & Tourism*, Vol. 23, pp.241–264, https://doi.org/10.1007/s40558-021-00196-4.

Poria, S., Chaturvedi, I., Cambria, E. and Hussain, A. (2016) 'Convolutional MKL based multimodal emotion recognition and sentiment analysis', *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp.439–448.

Shi, J., Li, W., Bai, Q., Yang, Y. and Jiang, J. (2023) 'Soft prompt enhanced joint learning for cross-domain aspect-based sentiment analysis', *Intell. Syst. Appl.*, Vol. 20, p.200292, https://doi.org/10.1016/j.iswa.2023.200292.

Song, Y., Wang, J., Jiang, T., Liu, Z. and Rao, Y. (2019) 'Attentional encoder network for targeted sentiment classification', *International Conference on Artificial Neural Networks*.

Tang, D., Qin, B. and Liu, T. (2016) 'Aspect level sentiment classification with deep memory network', *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, pp.214–224.

Tang, D., Qin, B., Feng, X. and Liu, T. (2015) 'Effective LSTMs for target-dependent sentiment classification', *International Conference on Computational Linguistics*.

Truong, Q. and Lauw, H. (2019) 'VistaNet: visual aspect attention network for multimodal sentiment analysis', *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, Hawaii, USA, pp.305–312.

Wang, Y., He, J., Wang, D., Wang, Q., Wan, B. and Luo, X. (2023) 'Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis', *Neurocomputing*, Vol. 572, p.127181, https://doi.org/10.1016/j.neucom.2023.127181.

Wu, C., Xiong, Q., Gao, M., Li, Q., Yu, Y. and Wang, K. (2020) 'A relative position attention network for aspect-based sentiment analysis', *Knowledge and Information Systems*, Vol. 63, pp.333–347, https://doi.org/10.1007/s10115-020-01512-w.

Xiang, R., Li, Z. and Sun, P. (2023) 'Research on sentiment analysis of scenic area comments based on RoBerta-BiGRU-attention--taking shenyang as an example', *Hans Journal of Data Mining*, Vol. 13, No. 4, pp.312–326.

Xu, J., Huang, F., Zhang, X., Wang, S., Li, C., Li, Z. and He, Y. (2019a) 'Visual-textual sentiment classification with bi-directional multi-level attention networks', *Knowl. Based Syst*., Vol. 178, pp.61–73, https://doi.org/10.1016/j.knosys.2019.04.018.

Xu, N., Mao, W. and Chen, G. (2019b) 'Multi-interactive memory network for aspect based multimodal sentiment analysis', *AAAI Conference on Artificial Intelligence*.

Xu, Q., Zhu, L., Dai, T. and Yan, C. (2020) 'Aspect-based sentiment classification with multi-attention network', *Neurocomputing*, Vol. 388, pp.135–143.

You, Q., Jin, H. and Luo, J. (2017) 'Visual sentiment analysis by attending on local image regions', *Proceedings of the AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, pp.231–237.

Yu, J., Jiang, J. (2019) 'Adapting BERT for Target-oriented multimodal sentiment classification', *Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp.1–7.

Zhang, J., Wu, X., Huang, C. (2023a) 'AdaMoW: multimodal sentiment analysis based on adaptive modality-specific weight fusion network', *IEEE Access*, Vol. 11, pp.48410–48420.

Zhang, K., Wang, S. and Yu, Y. (2023b) 'A TBGAV-based image-text multimodal sentiment analysis method for tourism reviews', *Int. J. Inf. Technol. Web Eng.*, Vol. 18, No. 1, pp.1–17.

Zheng, X., Huang, J., Wu, J., Sun, S. and Wang, S. (2023) 'Emerging trends in online reviews research in hospitality and tourism: a scientometric update (2000–2020)', *Tourism Management Perspectives*, Vol. 47, p.101105.

Zhu, D., Jing, R., Guo, Q., Zhang, D. and Wan, F. (2024) 'Sentiment analysis of tourism review text combined with bert-bilstm and attention mechanism', *J. Comput. Methods Sci. Eng*., Vol. 24, No. 3, pp.1605–1615.