The following publication Yang T, Hsu CH (2026), "An AI-driven framework for continuous tourist sentiment scoring using longitudinal and group-level insights with pre-trained language models (RoBERTa-CSS)". Tourism Review, Vol. 81 No. 1 pp. 167–187 is published by Emerald and is available at https://doi.org/10.1108/TR-05-2025-0550.

An AI-Driven Framework for Continuous Tourist Sentiment Scoring Using Longitudinal and Group-Level Insights with Pre-Trained Language Models (RoBERTa-CSS)

Abstract:

Purpose

Tourist sentiment is typically measured as discrete categories (e.g., positive, neutral, negative) through lexicon-based or machine-learning-based approaches in extant studies. However, neuroscience and physiology scholars have argued that sentiments are continuous in nature. Treating sentiment as a categorical state may result in an overly simplified understanding of tourists' sentiments, ultimately hindering the tourism industry's ability to derive precise and actionable insights.

Design/methodology/approach

This paper proposed a tool, RoBERTa-CSS (RoBERTa-based Continuous Sentiment Scoring), to calculate tourists' continuous sentiment scores based on the pre-trained language model RoBERTa. The structure of RoBERTa is refined by adding a fully-connected neural network layer so that continuous sentiment scores can be predicted. Using Chinese online reviews of a hotel group from multiple travel platforms, 3,500 sentences segmented from 1,000 randomly selected reviews were manually annotated to evaluate the proposed approach.

Findings

The comparison with the state-of-the-art open-source packages, deep learning models, pre-trained language models, and generative AIs on multiple evaluation metrics demonstrated the superiority of the proposed RoBERTa-CSS. The method was also validated on an English dataset, showing good performance. Several empirical analyses, including individual-level sentiment flow analysis, group-level sentiment distribution, and longitudinal analysis, were performed using the full dataset, and the results further showcased the edge of RoBERTa-CSS, compared to extant polarity categorization-oriented sentiment analysis methods.

Originality

This study expanded the analytical ability beyond simple categorization to facilitate

understanding of the complexity and diversity of human sentiment based on an improved pre-trained language model. The relevance of this paper for tourism practitioners, destination management organizations, and online travel platforms is discussed.

Keywords: Tourist sentiment; Continuous sentiment; Pre-trained language model; Big data analysis; RoBERTa-CSS

1. Introduction

Sentiment, a consistent evaluative state that combines cognitive and emotional processes, has emerged as a critical area of study across various disciplines due to its profound influence on individual decision-making, satisfaction, and long-term behaviors (Fu *et al.*, 2019; Hao *et al.*, 2020). In psychology, sentiments are considered enduring emotional dispositions that lead to responses toward individuals or objects in line with one's values (Cattell, 1940; Gervais and Fessler, 2017; Shand, 1922). Gervais and Fessler (2017) proposed the Attitude-Scenario-Emotion (ASE) Model, indicating that sentiment is a functional network of diverse basic emotions moderated across situations by attitudinal representations toward objects or individuals. They posited that attitudinal representations are psychological-level cognitions, evaluations, and affective tendencies exhibited by individuals, which moderate the triggering of their basic emotions in response to different scenarios. These discrete emotions and core attitudes interact and combine to form a functional network—referred to as sentiment—which maintains stability by having attitude moderate the relationship between scenarios and emotions. The stability of sentiment makes it a powerful predictor of behaviors, particularly in contexts such as tourism, where it plays a pivotal role in guiding tourist experiences and satisfaction (Mehraliyev *et al.*, 2022; Wu *et al.*, 2025). Hsu *et al.* (2016) conceptualized sentiment in tourism as "(people's) overall perceptions, views, and emotional dispositions underlying their responses to (the objects)" (p. 1). In this regard, understanding tourist sentiment is essential for improving service quality, destination branding, and operational decision-making in the tourism industry (Calderón-Fajardo *et al.*, 2024).

Given its significance, many studies have been conducted using and analyzing tourist sentiment based on user-generated content (e.g., posts on social media and online travel reviews) (Albayrak *et al.*, 2024; Wei *et al.*, 2023; Yang *et al.*, 2025). Typically, sentiment analysis classifies the sentiment (or valence) as positive, neutral, or negative (Mehraliyev *et al.*, 2020; Qiao *et al.*, 2022) based on the tone of the text. Commonly used sentiment analysis methods can be divided into two approaches: lexicon-based and machine learning-based. Lexicon-based methods rely on predefined sentiment dictionaries to assign discrete scores to textual data, offering simplicity and interpretability (Bagherzadeh *et al.*, 2021; Mehraliyev *et al.*, 2022).

Machine learning-based methods, including traditional algorithms like Support Vector Machines (SVM) (Costa *et al.*, 2019) and advanced deep learning models like Convolutional Neural Network (CNN) (Yang *et al.*, 2024), offer superior accuracy by learning and incorporating contextual information in the text. Based on deep learning algorithms and a massive corpus of text, some promising pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT), have emerged in recent years, which further enhance the understanding of contextual and situational information, thus leading to higher accuracy in sentiment analysis. For instance, Wu, Chen, *et al.* (2024) proposed utilizing BERT to categorize restaurant customers' sentiments into different categories (i.e., negative, neutral, and positive), and their results demonstrated greater precision than lexicon-based methods and traditional machine learning models.

The ASE Model suggests that sentiment is a complex outcome of the interaction of attitude, scenario, and emotion (Gervais and Fessler, 2017); thus, classifying sentiment as positive, neutral, and negative categories may oversimplify it (Gaspar *et al.*, 2016). The affect dimensionality theory (Russell, 1980) posits that affect can be characterized by continuous variations across multiple dimensions, including valence, which spans from extremely negative to extremely positive, capturing the nuanced and gradual nature of sentimental experiences. Barrett (2006) further supported this view, demonstrating that sentiment arises from the combination of core affects within specific contexts and sentimental responses are closely linked to continuous changes in brain activity and bodily states. The network of sentiment comprises multiple discrete emotions and core attitudes, with attitudes moderating the arousal of emotions in response to scenarios, the dispositions of which can range in value from low to high (Gervais and Fessler, 2017). In this regard, sentiment, like weight and length, should be a continuous variable, taking an uncountable infinite set of values. Unlike categorical variables, the continuous sentiment could be expressed with nuances such as "80% positive" or "30% negative", reflecting more granular evaluative states.

Current sentiment analysis typically categorizes tourist sentiment into discrete categories, such as -1, -0.5, and 0.5 in lexicon-based methods and negative and positive in deep learning-based methods. Given the rich information contained in sentiments, there is an urgent need to develop simple quantitative methods that can comprehensively describe people's sentimental

experiences in specific contexts (Galesic, 2017; Haslam, 2017). Moreover, lexicon-based methods suffer from lower accuracy (Yang *et al.*, 2024) due to the inability to capture contextual information of text. As a result, there is an increasing demand for a deep learning- or pre-trained language model-based sentiment analysis approach that can capture contextual information, facilitating more accurate and actionable insights for the tourism industry. Thus, the research question is, how can a continuous sentiment analysis method based on a pre-trained language model be developed?

To answer the research question, we proposed a pre-trained language model-based approach (RoBERTa-based Continuous Sentiment Scoring, RoBERTa-CSS) to calculate the continuous sentiment scores of tourists. Specifically, we first crawled 31,581 Chinese customer online reviews of a hotel brand in Hong Kong from multiple online platforms, including Trip.com, Booking.com, and TripAdvisor.com. Then, we constructed RoBERTa-CSS by improving the original model structure of Robustly Optimized BERT Pre-training Approach (RoBERTa) to produce continuous sentiment scores. To train and validate the proposed approach, we randomly selected 1,000 online reviews and recruited three tourism researchers to annotate these reviews at the sentence level. Based on the annotated dataset, we trained the proposed RoBERTa-CSS and compared it with several prevalent models, including two open-source packages (SnowNLP and Paddle), two deep learning models (CNN and LSTM), three pre-trained language models (ERINE, XLNet, and BERT), and two state-of-the-art generative AI tools (ChatGPT-4o and Qwen-long). The comparison demonstrated encouraging results of the newly proposed model.

This study overcomes the constraints of traditional sentiment analysis, which relies on discrete categorization. That is, the model treats sentiment as a continuous state instead of simply in categories of positive, neutral, and negative. This innovation pushes the conceptual boundary of the current sentiment analysis landscape. The proposed RoBERTa-CSS also advanced the methodological base of tourism research by providing a more accurate means of quantifying tourist sentiment on a continuous scale.

2. Literature review

2.1. Customer/tourist sentiment

In the psychology literature, sentiment is a concept that has garnered widespread attention due to its relative stability and impact on long-term judgments and behaviors (Naar, 2013). Cattell (1940) defined sentiment as "an acquired and relatively permanent major neuropsychic disposition" (p. 16). However, over the past two decades, psychology literature has primarily focused on the application of natural language processing or machine learning techniques in sentiment analysis tasks (Levis *et al.*, 2021; Oscar *et al.*, 2017; Zainal *et al.*, 2025), instead of extending the understanding or conceptualization of sentiment in a contemporary context. The ASE Model is one of the recent attempts to uncover the structure of sentiment, which defines sentiment as a functional network structured around a stable core of attitudes and diverse emotions across contexts, characterized by attitudes moderating the triggering of emotions by scenarios (Gervais and Fessler, 2017). As a response to external stimuli or physiological signals, sentiment is considered an enduring emotional disposition, as it is formed through sustained perceiving of a particular object and develops over time, and is associated with emotion and attitude (Munezero *et al.*, 2014; Yang and Hsu, 2025).

Nevertheless, sentiment is not a linear accumulation of attitudes or emotions. Sentiment is a system of multiple emotional dispositions (Shand, 1922) and may manifest varying discrete emotions in different contexts (Gervais and Fessler, 2017). Thus, emotion is fundamental, biological, and transient, while sentiment is complex, stable, and involves cognitions and judgment. While emotion depends on specific situations, sentiment can adjust to different circumstances and maintain its stability (Gervais and Fessler, 2017). The relative stability of sentiment arises not from the immutability of attitudes or emotions, but from consistent emotional responses to changing situations (Shand, 1922). Further, attitude was defined by Cattell (1940) as "an acquired neuropsychic disposition…as part of the purposive plan of some larger sentiment or complex" (p. 16). Chen *et al.* (2021) provided empirical evidence by measuring resident sentiment with cognitive and affective attitudes and other constructs. Based on the ASE Model (Gervais and Fessler, 2017), sentiment is more enduring and difficult to change by a single event, while highly informative events can directly alter attitude.

In tourism, researchers have emphasized the functional significance of sentiment in guiding individual decision-making and satisfaction (Mehraliyev *et al.*, 2022; Wu *et al.*, 2025). Thus, tourist/customer sentiment has received significant attention due to its potential to enhance customer experience, facilitate destination branding, and support operational decisions (Calderón-Fajardo *et al.*, 2024; Wu and Yang, 2023; Wu and Zhao, 2023). Sentiment-based insights have been used to assess service quality, identify customer preferences, and optimize marketing strategies in regional and global tourism markets (Liu, Huang, *et al.*, 2019; Wu *et al.*, 2023, 2025). For example, tourist sentiment enables Destination Management Organizations (DMOs) to assess public perceptions of attractions and adjust promotions accordingly, thereby strengthening destination image and competitiveness (Borrajo-Millán *et al.*, 2021). Additionally, research explores how sentiment data can inform personalized hospitality services, particularly through the integration of sentiment mapping into customer experience strategies (Rita *et al.*, 2023). Assessing the competitiveness of hospitality businesses based on customer sentiment is another prevalent application (Wu, Chen, *et al.*, 2024; Wu, Zhao, *et al.*, 2024). For example, Wu, Chen, *et al.* (2024) utilized customer sentiment to measure restaurant competitiveness by comparing it with that of one of the restaurant's competitors.

2.2. Sentiment analysis methods in tourism research

Given the importance of sentiment in understanding tourist/customer behaviors and experience and thus enhancing business performance, numerous tourism studies have conducted tourist sentiment analysis based on user-generated content (e.g., online reviews and social media posts) (Wu *et al.*, 2025; Wu and Yang, 2023). As for methods used, lexicon-based (Bagherzadeh *et al.*, 2021; Mahmoud *et al.*, 2025; Mehraliyev *et al.*, 2022) and machine learning-based (Luo and Xu, 2021; Yang *et al.*, 2024) methods have been the dominant approaches.

Lexicon-based sentiment analysis is a rule-based approach to assess and categorize sentiment in textual data by utilizing predefined sentiment lexicons—dictionaries that associate words or phrases with sentiment scores (e.g., -1, -0.5, 0.5, and 1) (Bagherzadeh *et al.*, 2021; Mehraliyev *et al.*, 2022). It has been widely applied in tourism research to assess customer

feedback. Researchers usually utilize general sentiment lexicons, such as SentiWordNet, to conduct analysis (León *et al.*, 2025; Mahmoud *et al.*, 2025; Qiao *et al.*, 2022). Similarly, scholars have used general open-source lexicon-based analysis tools to calculate customer/tourist sentiment; such tools include Linguistic Inquiry and Word Count (Shin and Nicolau, 2022) and Valence Aware Dictionary and sEntiment Reasoner (Varder) (León *et al.*, 2025). Despite their simplicity, they are unable to effectively capture domain-specific terminologies and contextual sentiment variations unique to tourism (Bagherzadeh *et al.*, 2021). To address this shortcoming, several studies have focused on developing domain-specific sentiment dictionaries that reflect industry-specific terms and customer concerns (Liu *et al.*, 2022; Tetzlaff *et al.*, 2019). For instance, Liu *et al.* (2022) incorporated a restaurant-domain-specific sentiment lexicon to enhance classification accuracy in hospitality reviews, while Tetzlaff *et al.* (2019) employed LASSO regression to build a tailored sentiment dictionary for hotel reviews.

Machine learning-based sentiment analysis is the use of machine learning algorithms/models to analyze sentiment, typically classifying sentiment into positive, negative, and neutral categories (Wu, Chen, *et al.*, 2024; Yang *et al.*, 2024). Traditional machine learning approaches include Support Vector Machines (SVM) (Costa *et al.*, 2019), Naïve Bayes, and Random Forest (Yang *et al.*, 2024). Researchers typically extract key features of text as input for machine learning models. For example, Yin and Jung (2024) first extracted the features of online reviews based on Term Frequency–Inverse Document Frequency (TF-IDF) and then applied SVM to classify tourists' sentiment into positive and negative polarities. Despite achieving better accuracy than lexicon-based methods, machine learning-based methods also suffer from poor contextual understanding. With the development of neural networks, deep learning techniques have shown superior performance in sentiment classification by capturing sequential and contextual dependencies in textual data (Luo and Xu, 2021; Yang *et al.*, 2024). Commonly used techniques include Convolutional Neural Network (CNN) (Yang *et al.*, 2024) and Long Short-Term Memory (LSTM) network (Wu, Zhong, *et al.*, 2024). More recently, transformer-based models like BERT and its variants (e.g., RoBERTa) have demonstrated higher accuracy by leveraging self-attention mechanisms to effectively encode contextual relationships in reviews (Liu and Hu, 2023; Wu, Chen, *et al.*, 2024).

In summary, regardless of which approach is used to construct the lexicon, lexicon-based sentiment analysis essentially calculates discrete sentiment values for the visitors/consumers, relying on the discrete sentiment scores of words in the lexicon. Moreover, its inability to understand the deeper semantic connotations of the text leads to lower accuracy of the sentiment computation (Yang *et al.*, 2024). As for machine/deep learning-based methods, they typically classify sentiments into discrete categories, ignoring the continuous disposition of sentiment. Affect dimensionality theory indicates that affect can be described by successive changes in multiple dimensions, which range from extremely negative to extremely positive, reflecting the gradual movement and complexity of this concept (Russell, 1980). In addition, in practical applications, continuous dimensions can more accurately capture subtle changes in sentiment; therefore, individual tourist sentiment should be viewed as a continuous spectrum rather than a discrete categorization. Thus, an advanced and accurate approach to calculating tourists' continuous sentiment needs to be developed.

2.3. BERT-based sentiment analysis

BERT is a pre-trained language model based on the Transformer encoder proposed in 2018, whose core innovation lies in learning general-purpose language representations through bi-directional context modeling and masked language modeling (MLM) tasks (Devlin *et al.*, 2019). MLM in this context refers to randomly masking a portion of words (typically 15%) in the input text and then allowing the model to predict these masked original words, an approach that forces the model to understand the full contextual information to accurately perform the prediction task (Bao *et al.*, 2020; Cui *et al.*, 2021). The model employs a multi-layer Transformer encoder stacking structure with fused inputs of word embeddings, positional embeddings, and segmental embeddings. Then it is pre-trained on a large-scale unlabeled corpus using MLM and the Next Sentence Prediction (NSP) task, a self-supervised learning approach that allows it to capture rich semantic and syntactic information (Devlin *et al.*, 2019). In the application phase (e.g., sentiment analysis), BERT can adapt to downstream tasks by simple fine-tuning. For example, extant studies have fine-tuned BERT for sentiment categorization tasks in various contexts, including restaurant competitiveness analysis (Wu, Chen, *et al.*, 2024), customer satisfaction analysis (Yang *et al.*, 2024), and business survival

prediction (Li *et al.*, 2023).

Despite the broad application of BERT, its shortcomings are also salient, so several variants (e.g., RoBERTa by Liu, Ott, *et al.* (2019) and SpanBERT by Joshi *et al.* (2020)) have been developed. As an important enhanced version of BERT, RoBERTa significantly improved the model performance through systematic optimization of the training strategy (Liu, Ott, *et al.*, 2019). For a clearer understanding of its model structure, Figure 1 presents the pre-training process of RoBERTa. Specifically, the training process of RoBERTa is based on the Transformer encoder (i.e., Trm in Figure 1) architecture and uses a dynamic masked language model task (e.g., "Masked Sentence A" and "Masked Sentence B") for pre-training: the model first represents the input text as tokens, then randomly selects a few tokens to be dynamically masked (re-selecting the masking position each time it is trained), and finally, through a multilayer Transformer encoder learning to predict the masked original tokens (Liu, Ott, *et al.*, 2019). Unlike BERT, RoBERTa removes the Next Sentence Prediction (NSP) task, employs larger batch sizes and longer training cycles, and extends the amount of training data to 160GB to enable the model to learn linguistic representations more comprehensively (Liu, Ott, *et al.*, 2019).

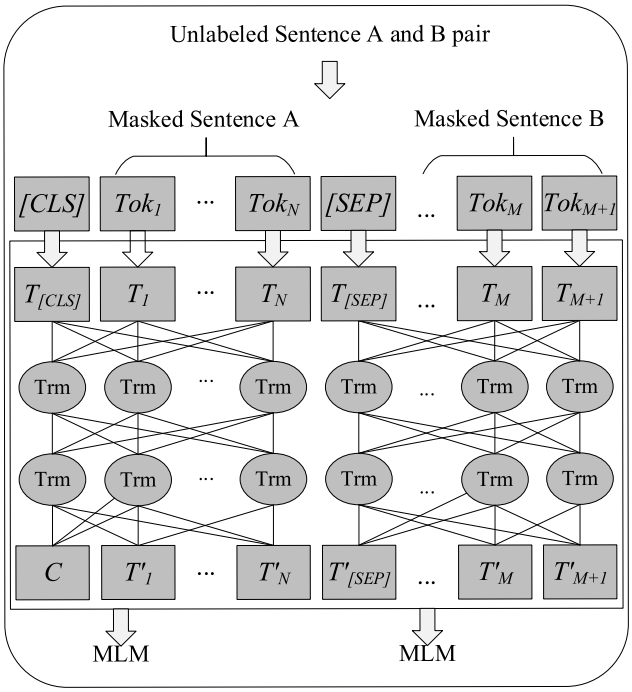


Figure 1 Process of pre-training RoBERTa (Adapted from Devlin et al. (2019))

Note: (1) CLS is the abbreviation of classification, located at the beginning of the text sequence; (2) Tok is the abbreviation of token; (3) SEP is the abbreviation of separator, separating sentences.

For the sentiment analysis task, these improvements of RoBERTa bring significant advantages. For instance, the larger training data size enables the model to capture richer sentiment expressions. In particular, the dynamic masking strategy enhances the model's robust understanding of the sentiment vocabulary, and the removal of the NSP task allows the model to focus more on the extraction of sentiment features from the text itself. In addition, RoBERTa's understanding of negatives and sentiment modifiers is also more accurate, demonstrating its unique advantages in complex tasks such as sentiment intensity prediction. However, extant tourism research based on RoBERTa has typically applied the model in sentiment categorization (e.g., Vogklis & Gkritzali, 2024) rather than constructing methods that can predict continuous sentiment scores. Thus, the current study aims to fill this gap.

3. Methodology development

To address the research gaps identified from the literature, this section aims to propose a method to calculate continuous sentiment scores based on the comprehensive methodology framework shown in Figure 2. It mainly includes four stages: (1) Data collection and pre-processing, (2) Data annotation, (3) Constructing RoBERTa-CSS, and (4) Empirical analysis using RoBERTa-CSS. In the following subsections, we will detail the first three stages, and then in the Results section, we further report the model construction outcomes and empirical analysis findings.

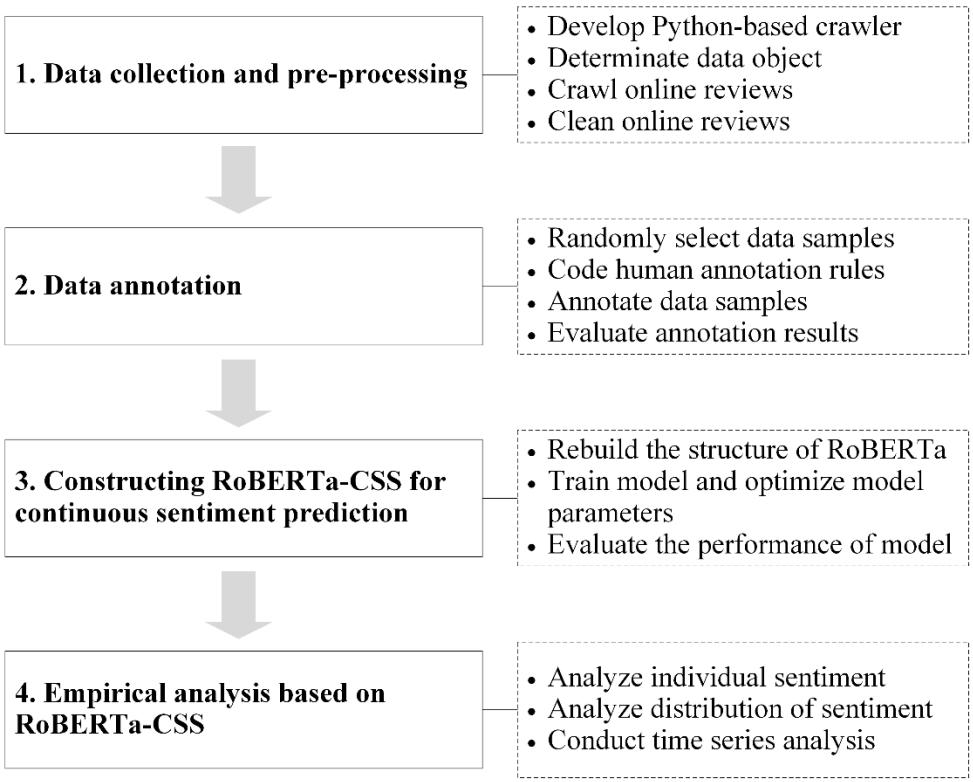


Figure 2. Methodology framework

3.1. Data collection and pre-processing

To avoid data sampling bias, we diversified our data sources by using multiple online platforms, including Trip.com, Booking.com, and TripAdvisor.com. These platforms have been commonly used as research data sources (Guo *et al.*, 2024; Mariani and Borghi, 2021; Peng *et al.*, 2018). We selected one hotel brand in Hong Kong, which has 12 hotels in operation, as our research object. After manually recording the URLs of these hotels on the three online platforms, we subsequently harnessed Octoparse (also called Octopus or bazhuayu) (Shenzhen Shukuo Information Technology Co. Ltd, 2013), a widely used data crawler (Yang *et al.*, 2024; Wu *et al.*, 2023), to scrape online reviews of these hotels. After removing duplicate reviews and those without review text, we obtained 31,581 pieces of online reviews in Chinese, including multiple fields such as review text, overall rating, and posting date.

3.2. Data annotation

Since this research aimed to propose a pre-trained language model-based approach to derive continuous sentiment scores of customer/tourist reviews, a standard dataset with sentiment score labels should be developed. Of the 31,581 online reviews collected, 1,000 were randomly selected to construct the annotated dataset. Different from prior studies that focused

on classifying reviews into different sentiment polarities, our work emphasized the continuity of sentiment. In this regard, the annotation process of continuous sentiment scores of review text was purposely designed. People often include multiple sentiments in an entire paragraph, and sentences are regarded as units of natural language that express complete thoughts or sentiments. Considering the number of hotel aspects is relatively few (e.g., TripAdvisor summarizes 6 hotel aspects) and people usually write long posts (e.g., 8 or 10 sentences per post), sentence-level analysis can offer more details than aspect-level analysis. In addition, sentence-level sentiment analysis does not rely on the accuracy of aspect identification and can capture the overall sentiment of sentences. Thus, we segmented these 1,000 reviews into 18,164 sentences and saved them in an Excel file. From sentences of moderate length (i.e., 10 to 50 Chinese characters), we randomly selected 3,500 sentences for annotation. Then, annotators with tourism knowledge were recruited and asked to annotate the sentiment of the sentences independently using a continuous scale, following the rules below:

(1) Use a continuous scale from -5.0 to 5.0, in 0.5 increments, to annotate the sentiment of each text sample.

(2) Consider the intensity of the sentiment expressed in the text.

(3) Consider the context in which the text was written, such as sarcastic and ironic tones.

(4) If the text mentions specific entities (e.g., people, organizations, products) or aspects (e.g., features, attributes), consider the sentiment towards each entity or aspect separately, then average them on the sentence level.

(5) If the sentiment is ambiguous or uncertain, annotate with a value close to 0 (neutral).

(6) If a text sample is invalid or meaningless, annotate it with "F".

After annotating the dataset at the sentence level, we removed those sentences with an "F" label and obtained 3,449 of them with continuous sentiment scores. The labeling of a small range of values (e.g., 0 to 1) can make the gradient update smoother and reduce the risk of gradient explosion or disappearance (Pascanu *et al.*, 2013), making the model converge faster. Thus, we normalized the annotation results of the three individuals to [0,1] separately and calculated the average of their annotations as the final annotation results at the sentence level.

Based on the proposed annotation rules, we obtained annotated results from three research assistants who worked independently. To validate the annotated results, we adopted the Pearson

correlation coefficient (Fong *et al.*, 2013) to assess the consistency among the results by different annotators. That is, when the annotation results of the three annotators are similar (consistent), the quality of the annotation is deemed high. Essentially, this principle is similar to testing the labeling quality of categorical variables. Accordingly, we calculated the Pearson correlation coefficients for any two annotation results, and they were 0.8330, 0.8545, and 0.8526, which indicate a high annotation reliability. Subsequently, we calculated the average of these three individual ratings as the final sentiment score.

3.3. Constructing RoBERTa-CSS

3.3.1. Algorithmic structure of RoBERTa-CSS

Figure 3 shows the core algorithmic structure of the proposed RoBERTa-CSS, including five coding steps (see the codes in the online Appendix A). The grey part corresponds to the RoBERTa model in Figure 1, and the blue part is the innovation part constructed in this paper.

- Step 1 transforms the raw text into a batch tensor format that can be processed by the model and enables automatic batching and random disruption of the data. We use one sentence as an example in Figure 3 to illustrate the algorithmic process.
- Step 2 converts the raw text into a sequence of numeric IDs required by the model, including Token Embeddings, Segment Embeddings, and Position Embeddings, utilizing the pre-trained RoBERTa model in Figure 1, and standardizes the input length to support batch processing.
- Step 3 obtains the [CLS] token, which is designed to carry the semantic information of the entire sequence during pre-training and serves as an input feature for the subsequent regression layer, which is able to preserve the key semantic features.
- Step 4 maps the 768 semantic feature dimensions (obtained from the original RoBERTa structure, Liu, Ott, *et al.* (2019)) to the regression function (i.e., the Regressor Layer in Figure 3).
- Step 5 aims to use *MinMaxScaler*, a standardization method in Python, to scale the predicted sentiment scores into the interval [0, 1], following Equation (1), where Y is the originally predicted sentiment score set and $\tilde{y}$ is the standardized predicted sentiment score y. This operation helps normalize the sentiment scores in different ranges to make them comparable.

$$\tilde{y} = \frac{y - \min(Y)}{\max(Y) - \min(Y)} \tag{1}$$

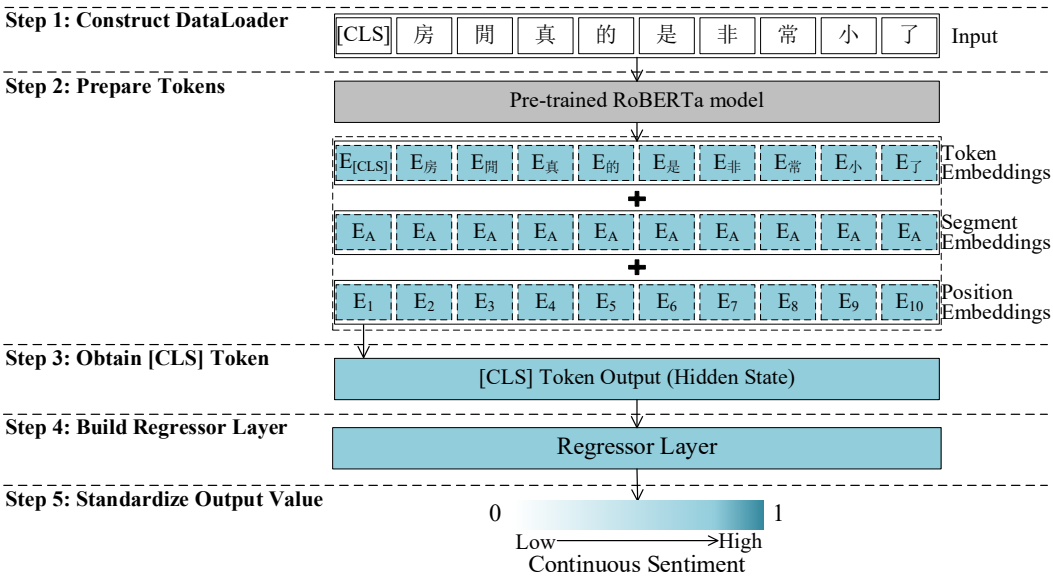


Figure 3 Architecture of the proposed RoBERTa-CSS

Note: Grey layer comes from original Ro-BERTa; Blue layers are newly proposed in this work.

Overall, the current approach inherited RoBERTa's ability to extract contextual features from text. That is, the proposed approach can capture the gradual change and context-dependent sentiment embedded in text by incorporating all semantic information of the sentences. In addition, by changing the output layer to a regression function, RoBERTa-CSS can produce continuous sentiment values between 0 and 1, thus quantifying the intensity of the sentiment more accurately and in more detail than performing simple categorization only. Therefore, using RoBERTa-CSS to compute continuous sentiment scores is both theoretically and practically sound and efficient.

3.3.2. Process of fine-tuning and optimizing RoBERTa-CSS

After constructing the RoBERTa-CSS based on the original RoBERTa, the model was fine-tuned to optimize its performance using the annotated dataset. The process of fine-tuning began with loading the pre-trained "hfl/chinese-roberta-wwm-ext" (Cui *et al.*, 2021) model as the infrastructure and adding a linear regression layer on top of it to form a complete regression model. The model utilizes the [CLS]-labeled hidden states output as regression inputs and predicts continuous sentiment values through a fully connected layer, as shown in Equations (2)-(4):

$$O = RoBERTa(S) \tag{2}$$

$$h_{[CLS]} = O[:,\ 0,\ :] \tag{3}$$

$$y = W^T h_{[CLS]} + b \tag{4}$$

where O is the output of the last hidden state layer of the model. In particular, we take the output corresponding to the [CLS] token as the representation of the whole sentence, as shown in Equation (2). Then, $h_{[CLS]}$ is fed into a linear layer for sentiment score prediction, as shown in Equation (3), where W and b are the weights and biases of the linear layer, respectively, and y is the predicted sentiment score. During the training process, an end-to-end fine-tuning approach was used to simultaneously update the parameters of the RoBERTa-CSS backbone network and the added Regressor Layer to adapt the model to specific sentiment analysis tasks.

For the optimization process, we used the AdamW optimizer (Alammary, 2025) in conjunction with the learning rate linear warmup scheduling, with the initial learning rate set to 2×10^{-5}. The learning rate scheduling formula is:

$$\eta_t = \eta_{\max} \times \min(\frac{t}{T_{warmup}}, 1) \tag{5}$$

where T_{warmup}=1000. The training objective is to minimize the loss value of the model, as shown in Equation (6),

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

where y_i is the true label, $\hat{y}_i$ is the predicted value, and n is the sample size. Because the loss value is computed after squaring, Equation (6) is very sensitive to large prediction errors and can significantly amplify the effect of these errors. This property makes *Loss* a proper optimization objective. In order to enhance the training stability and prevent overfitting, an early-stop mechanism is implemented to monitor the validation set loss and terminate the training early when the validation loss no longer decreases in consecutive rounds.

3.3.3. Metrics for evaluating the proposed model

As the model developed measures sentiment using a continuous scale, the predicted results are continuous values. Therefore, prevalent evaluation metrics for model training and validation, such as *Accuracy* and *F1* scores (Wu, Chen, *et al.*, 2024), which are typically

designed for discrete labels, may not be appropriate to apply in our research context. Further, given the contextual characteristics of the proposed model, we selected two indicators, namely mean square error (*MSE*) and mean absolute error (*MAE*), to assess the performance of the model.

Specifically, *MSE* is a measure of the average of the squared error between the predicted value and the true labels. Equation (7) shows its calculation process. Notably, although the formulas for *MSE* and *Loss* (in Equation (6)) are essentially the same (both are mean square error), they are different because of the range of data and the purpose of the phase in which they are computed. Specifically, *Loss* is computed in real time for each batch during training and is used to optimize the model for back-propagation. *MSE* is computed at the end of the entire epoch for all samples in the training or validation set and is used to evaluate the model's performance. *MAE* measures the predictive performance of a model by calculating the average of the absolute values of the error between the predicted and true labels. Equation (8) shows its calculation process. In addition, consistent with existing research (Cheng *et al.*, 2024), we use the loss function to determine the training epochs.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{8}$$

4. Results

4.1. Comparison with extant methods

After training (see Appendix B for more details), our proposed model was compared with four types of existing methods. Specifically, among the multiple models that can produce continuous sentiment scores without using annotated datasets, open-source packages are freely available and generative AI tools are gaining popularity. SnowNLP (Li *et al.*, 2025) and Paddle (Wu, Zhao, *et al.*, 2024), two prevalent open-source Python packages, have been widely utilized in tourism studies. In addition, one of the most advanced generative AIs, ChatGPT, can be used to predict tourist sentiment. With appropriate prompting words, ChatGPT can produce continuous sentiment scores (Cheng *et al.*, 2024). Thus, we selected the most updated version of ChatGPT (as of this paper's writing), namely ChatGPT-4o, as a comparative model.

Additionally, because this study annotated Chinese reviews as the training dataset, we selected a popular Chinese generative AI tool, Qwen-long, developed by Alibaba, for comparison purposes. Moreover, considering the superior capability of deep learning models and pre-trained language models to understand natural language, we also incorporated traditional deep learning algorithms, CNN and Long-Short Term Memory network (LSTM), and existing pre-trained language models, including BERT, XLNet, and ERNIE, for comparison purposes. To enable pre-trained language models to score continuous sentiments, we added regressor layers to the original models, as we did in the process of refining RoBERTa.

After running the above models based on our annotated dataset (i.e., 3,500 sentences mentioned in Section 3.2), we evaluated the performance of all models, as shown in Table 1. Among all models, the open-source package SnowNLP performed the worst, obtaining the highest *MSE* value, indicating a large difference between the model's predicted and annotated values. It also achieved the highest *MAE*, reflecting a large average absolute difference between the model's predicted and annotated values. In addition, Paddle showed poor performance, with high *MSE* and *MAE* values. Similarly, deep learning-based models (i.e., CNN and LSTM) did not show good performance. In contrast, the pre-trained language models all achieved better results; and in particular, BERT performed the best among them. Moreover, the two generative AI tools presented good performance in terms of *MSE* and *MAE* values, notably lower than those of the open-source packages and deep learning models tested, but their cost stands out. For example, the unit cost of invoking ChatGPT-4o via Python is 2.5 USD per 1 million tokens, and to compute the continuous sentiment scores of 100,000 medium-length reviews (e.g., 150 Chinese characters, around 300 tokens), with 60 prompting words (i.e., 60 tokens), the cost of a single invocation would be about (300+60)* 2.5*100,000/1,000,000=91.25 USD. This cost is only for a single data analysis, and it would be much higher for medium-sized businesses conducting quarterly or monthly sentiment analysis.

Table 1 Comparison results

Type	**Model**	**MSE**	**MAE**	**Cost**
Open-source packages	SnowNLP	0.21	0.38	Free
	Paddle	0.06	0.20	Free
Deep learning models	CNN	0.07	0.23	Free
	LSTM	0.07	0.23	Free

Pre-trained language models	ERNIE + regressor layer	0.04	0.17	Free
	XLNet + regressor layer	0.02	0.09	Free
	BERT + regressor layer	0.01	0.08	Free
Generative AI tools	Qwen-long	0.01	0.07	$1 / 1M tokens
	ChatGPT-4o	0.01	0.07	$2.5 / 1M tokens
Proposed model	**RoBERTa-CSS**	**0.01**	**0.06**	Free

Note: (1) We also tested the effects of the proposed model on an English dataset; see the online Appendix C for details. (2) Since user-generated reviews often contain typographical errors, ambiguous statements, or incomplete sentences, which can challenge the model's predictive accuracy, we further tested the performance of the proposed model on a dataset with noise, as shown in Appendix D.

Encouragingly, the RoBERTa-CSS demonstrated the lowest *MSE* and *MAE* values (i.e., 0.01 and 0.06) among all models, indicating that the error between the predicted and annotated values of the model is small, which means the model's prediction accuracy is high. To enable pre-trained language models to calculate continuous sentiments, we utilized the same strategy (i.e., adding a regressor layer) as in the RoBERTa-CSS. Thus, these models, especially XLNet and BERT, also performed well. The results demonstrate that the strategy of adding a regressor layer to a pre-trained language model is an effective way to obtain continuous sentiment scores, and that RoBERTa-CSS performed the best. In addition, as usage cost might be a concern for researchers when conducting big data analysis, we further supplemented the cost information in the comparison experiment, as shown in the last column of Table 1. While many AI tools are open to the public, often AI tools limit the quota of free trials available to unpaid users. In other words, when individuals or businesses need to process large-scale datasets using AI tools, they often need to consider the cost involved. In this regard, the proposed model presents an advantage (i.e., free to use and high accuracy) in actual usage over other tools.

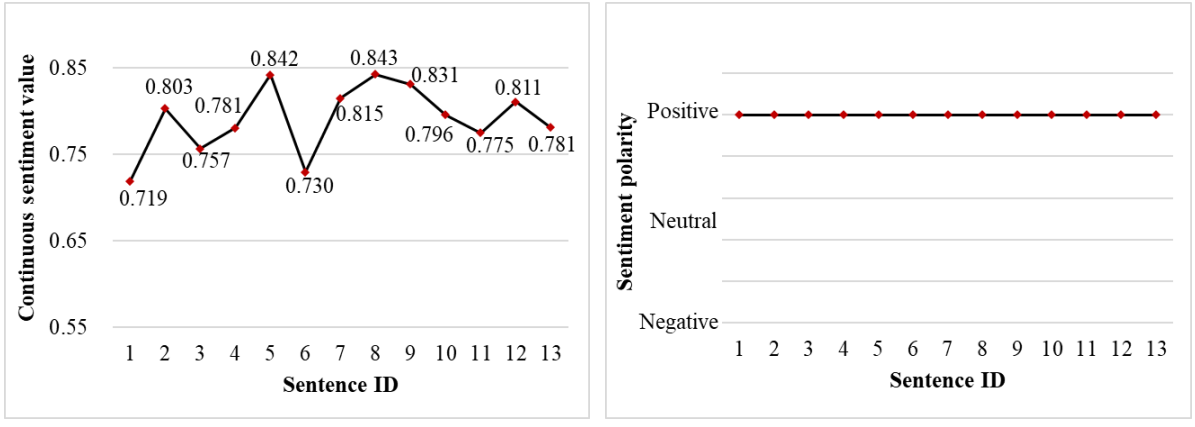
4.2. Empirical analysis using RoBERTa-CSS

This study highlighted the importance of understanding tourist sentiments on a continuous scale and thus developed a pre-trained language model-based approach to measure continuous sentiment scores. As an initial effort to demonstrate the necessity of obtaining sentiment value on a continuous scale, in this section, we illustrate the empirical applications of the proposed model, including individual-level sentiment flow analysis and group-level sentiment distribution analysis. We also conducted a group-level longitudinal analysis, which is reported in Appendix E.

(1) Individual-level sentiment flow analysis

While sentiments are stable in nature, it is difficult for humans to maintain sentiments as constant or to change dramatically from extremely positive to extremely negative. Thus, depicting details of individuals' sentiment flow can increase our understanding of the concept and more clearly visualize the concept in action. As discussed, extant tourism studies typically classified tourist sentiment into several categories (Wu, Chen, *et al.*, 2024; Yang *et al.*, 2024), ignoring the continuous nature of sentiment. Thus, we first compare differences between sentiment categories and continuous sentiment scores in understanding sentiment flow patterns. To obtain sentiment categories, we refer to the work of Wu, Chen, *et al.* (2024), using BERT to group text into three categories (i.e., positive, neutral, and negative). Meanwhile, utilizing our proposed model, we calculated the continuous sentiment values, ranging from 0 to 1.

Taking one online review as an example, we first segmented it into 13 sentences and then used two methods to calculate the sentiment value of each sentence according to the order of the sentence in the online review. Figure 4 illustrates the findings, where subfigure (a) represents the result of RoBERTa-CSS and (b) represents the result of BERT. Figure 4 (b) shows that the extant method categorized all sentences as positive. In contrast, Figure 4 (a) shows the continuous sentiment value of each sentence. Even though sentiments in all sentences are positive, we can observe the fluctuation of tourist sentiment in this online review, which wavers in the first few sentences and stabilizes at the end.



(a) Continuous sentiment scores (b) Discrete sentiment categories

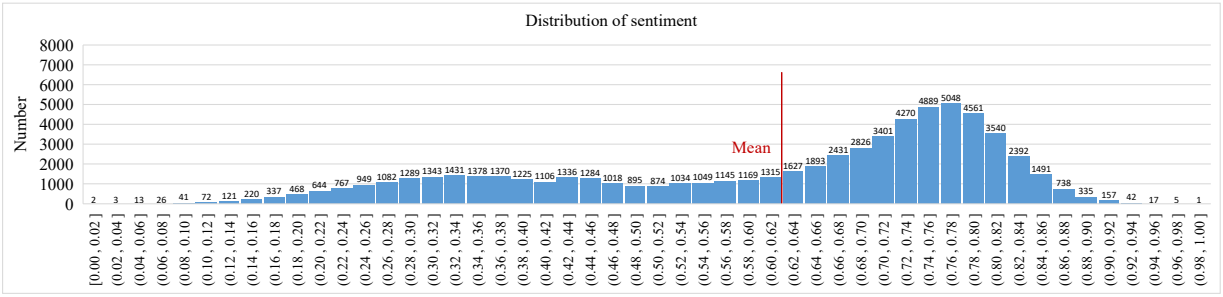
Figure 4 Results of individual sentiment flow analysis

Hotels' frontline departments, such as front office, housekeeping, and food & beverage, can benefit from the use of RoBERTa-CSS. As shown in Figure 4, RoBERTa-CSS can monitor

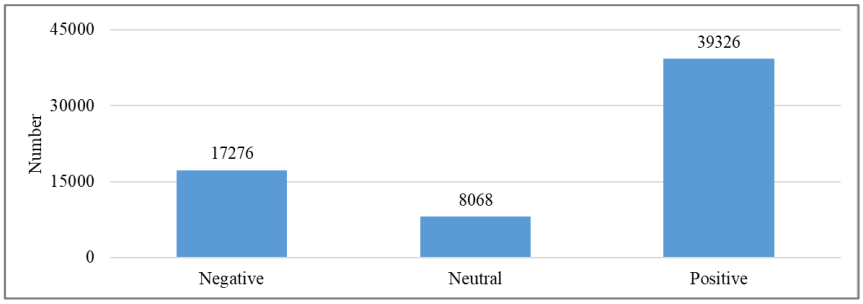
customers' individual-level continuous sentiments expressed in sequential sentences. Through alert of customer relative sentiments, these frontline departments are able to keep abreast of potential triggers of lower levels of consumer satisfaction and target remediation. For example, Sentences 1 and 6 in Figure 4 (a) could be the focus of managerial attention.

(2) Group-level sentiment distribution analysis

The group-level distribution of sentiment calculated by the proposed model using all 31,581 online reviews collected (i.e., 64,670 sentences) was compared with distribution calculated using an existing sentiment categorization-oriented method (Wu, Chen, *et al.*, 2024), as shown in Figure 5 (a) and (b), respectively. Compared with Figure 5 (b), Figure 5 (a) contains and reveals much more information. Specifically, the dataset includes 85 extremely negative sentences (value lower than 0.1) and 222 extremely positive ones (value higher than 0.9). Moreover, RoBERTa-CSS enabled the performance of statistical analysis, such as calculating the mean value, thus quantifying the overall level and nuanced variations of tourist sentiment. In contrast, Figure 5 (b) only reveals that 39,326 sentences are positive, 17,276 sentences are negative, and 8,068 sentences are neutral, which means most sentences are positive.



(a) Result of the proposed RoBERTa-CSS



(b) Result of extant sentiment classification model

Figure 5 Comparison of the distribution of sentiment

Note: The numerical interval of the horizontal axis in (a) is adjustable (currently 0.03), as our method can

compute continuous sentiment scores.

The general office and procurement department can leverage RoBERTa-CSS to achieve more rational resource allocation and service optimization based on continuous sentiments. Figure 5 (b), which displays the traditional sentiment categories, suggests that the resource allocation should focus on the 17,276 negative sentences. However, RoBERTa-CSS provides relevant departments with more precise decision-support evidence, i.e., slightly positive sentences also need attention in resource allocation.

5. Implications

5.1. Academic and methodological implications

Echoing Haslam's (2017) and Galesic's (2017) appeal of simple quantitative models to precisely describe sentiment, the current research built an AI-driven framework for continuous tourist sentiment scoring. From a neuroscientific and psychological perspective, human sentiment is continuous in nature, as the sentiment responses triggered by changes in brain activity and body states are continuous (Barrett, 2006). By revealing subtle variations in sentiments (i.e., positive sentiments can shift between mildly positive and intensely positive), the present study calls for tourism researchers' attention and adherence to the continuous nature of sentiments. Due to prior methodological limitations, sentiments have been examined in an oversimplified form. Existing studies usually use lexicon-based or machine/deep learning-based sentiment analysis (Fu *et al.*, 2019; Hao *et al.*, 2020; Wu, Chen, *et al.*, 2024); both of which quantify tourists' sentiment in discrete categories. Staying true to the conceptualization of sentiment enables us to go beyond describing sentiment with empirical data. This research overcomes the constraints of discrete categorization in traditional sentiment analysis methods and treats sentiment as a continuous spectrum. By providing a reliable tool for continuous sentiment measurement, we advance theoretical progress and benefit the subsequent development of sentiment research.

The proposed model introduced an advanced pre-trained language model-based approach, RoBERTa-CSS. Building on the structure of the original RoBERTa model, we added a linear neural network layer (i.e., Regressor Layer in Figure 3) after the Token Output Layer so that the proposed model can output the continuous values of sentiment. Thus, the current study

offers a new big data analytics method that outperforms two open-source packages (SnowNLP and Paddle), two deep learning models (CNN and LSTM), three pre-trained language models (ERINE, XLNet, and BERT), and two state-of-the-art generative AI tools (ChatGPT-4o and Qwen-long). As shown in Figure 4 (a), sentence-level continuous sentiment scores uncovered differences that cannot be recognized by traditional sentiment classification methods. The proposed method enabled a refined understanding of the subtleties in tourist sentiment. Further, as it is free to use, the method can facilitate more researchers to investigate sentiments using continuous scoring. Accordingly, the current study has methodological contributions to sentiment research by introducing a tool for recognizing nuanced sentiment at a low cost.

5.2. Practical implications

The model developed can more accurately capture the intricacies of tourist sentiments, reflecting that travel experience cannot be simply summarized as having an overall positive or negative sentiment. Traditional sentiment categorization methods are likely to overlook pertinent issues that require attention due to an overall "positive" sentiment label, as they are unable to identify different aspects having varied levels of sentimental states. As the tourism space becomes more competitive, destinations and businesses cannot satisfy all tourists with standardized marketing and service offerings. RoBERTa-CSS empowers tourism practitioners with evidence-based insights on tourist sentiments to develop targeted marketing, communication, and service strategies, thereby enhancing their competitiveness. For example, businesses are advised to prioritize resources to address the areas of concern expressed by tourists as negative, yet embedded in an overall positive online review. This would resolve issues causing potential dissatisfaction and negative e-word-of-mouth (eWOM). In addition, by analyzing the perceptions of tourists associated with different continuous sentiment values, businesses can locate the tipping point of tourist dissatisfaction. For example, tourists may not complain when the sentiment value of a particular aspect of the service is slightly negative, yet when the sentiment score reaches a certain level, negative eWOM begins to emerge.

As illustrated in Appendix D, through time-series analysis, RoBERTa-CSS can reveal the dynamic trend of customer sentiment, whereas traditional sentiment classification models have only been used to provide discrete sentiment labels. This ability to analyze continuous

sentimental values enables DMOs to detect changes in customer sentiment in a timelier and more accurate manner, especially during atypical periods (e.g., outbreak of an epidemic, issuance of a new policy, or after a major event). For example, DMOs can monitor any rapid increase in negative sentiments to understand continuous changes in tourists' sentiment values and take appropriate remedial measures at the destination or specific industry segment level.

RoBERTa-CSS refines a current model and offers advantages over other state-of-the-art models. It has not only demonstrated superior performance compared to the tested open-source packages, deep learning models, pre-trained language models, and generative AI tools, but also is cost-free to use. While many generative AI tools offer free trials, users or businesses are often required to pay for access to the services when handling large-scale datasets. Thus, RoBERTa-CSS is particularly advantageous for research organizations and enterprises that need to handle massive amounts of online reviews in the current big data era. The method developed requires no further training and would not change, which means there are no re-learning or adaptive functions. The model will be the same for everyone using it, and every time they use it, which ensures the transparency of the algorithm. Thus, this model addresses Hsu *et al.*'s (2024) concerns of users influencing future query results. It can be directly applied to various tourist datasets following a standard procedure. First, practitioners can collect tourist-generated textual data and organize the data by time. Then, data should be pre-processed by deleting noise, incomplete data, and meaningless data, and segmenting the retained data into sentences. After that, using the codes reported in the online Appendix A and the experimental environment and configuration shown in Table B1, businesses can efficiently analyze millions of user reviews on a regular desktop computer, thus better understand customer opinions and optimize products and services accordingly.

6. Conclusion, limitations, and future agenda

Emphasizing the importance of understanding tourist sentiment as a continuum, the current paper proposed a supervised learning method, i.e., RoBERTa-CSS, calculating tourists' continuous sentiment scores using labelled data in the training process. Although unsupervised methods, such as zero-shot learning and sentiment dictionary, are available, their accuracy is found to be quite low. Specifically, we built on the model structure of the original RoBERTa

by adding a linear fully connected layer after its attention encoder layers, based on which we predicted and produced continuous sentiment scores of tourist reviews. Regarding the training data, we employed a manual annotation task instead of utilizing Gen AI or other tools to automatically label data, because AI tools are found to have poor performance in tourism contexts (Saleh, 2025; Seyfi *et al.*, 2025; Statista, 2024). Due to variations in sentiment expressions of human languages, human annotators with domain knowledge can better label the data. To address potential annotator bias, we recruited three annotators to label the data independently, which may decrease adverse impacts. After finalizing the model's parameters, we compared it with extant prevalent models, including open-source packages and generative AI tools, and found encouraging results based on multiple metrics, including *MAE* and *MSE*. Even though the labelling process is time-consuming, future users of the proposed method do not need to label data again, because the model has been developed and is ready to use as is without further training. In addition, we demonstrated how continuous sentiment scores can benefit individual-level sentiment variation analysis and group-level analyses. Such empirical analyses further validated the superiority of the proposed model in revealing finer insights into tourist sentiment.

Although this study has made progress in continuous sentiment score prediction, there is still room for further work. First, the training data of the model is based on reviews in a single language—Chinese. Although RoBERTa-CSS performs well in analyzing both Chinese and English texts, its performance in multilingual and multicultural environments can be further improved and verified by including data in more languages (e.g., Spanish, Korean, and French) in future training and application on a global scale. Second, the current research focused on sentence-level continuous sentiment scores. Given the potential added value of finer-grained aspect-level sentiment analysis for understanding tourist sentiment, the application of RoBERTa-CSS can be extended to the context of aspect-level continuous sentiment scoring, as multiple aspects could be present in a sentence. In addition, automated annotation processes can enhance the efficiency of pre-trained language models, but existing AI tools have been found to perform poorly when applied in tourism contexts (Saleh, 2025; Seyfi *et al.*, 2025; Statista, 2024). Thus, the construction of accurate automated annotation AI tools is encouraged in future research.

References

Alammary, A.S. (2025), "Investigating the impact of pretraining corpora on the performance of Arabic BERT models", *The Journal of Supercomputing*, Springer, Vol. 81 No. 1, p. 187.

Albayrak, T., Dursun-Cengizci, A., Fong, L.H.N. and Caber, M. (2024), "The changing role of hotel attributes in destination competitiveness throughout a crisis", *International Journal of Contemporary Hospitality Management*, Vol. 36 No. 10, pp. 3264–3282, doi: 10.1108/IJCHM-06-2023-0779.

Bagherzadeh, S., Shokouhyar, S., Jahani, H. and Sigala, M. (2021), "A generalizable sentiment analysis method for creating a hotel dictionary: using big data on TripAdvisor hotel reviews", *Journal of Hospitality and Tourism Technology*, Emerald Publishing Limited, Vol. 12 No. 2, pp. 210–238, doi: 10.1108/jhtt-02-2020-0034.

Bao, H., Dong, L., Wei, F., Wang, W., Yang, N., Liu, X., Wang, Y., *et al.* (2020), "Unilmv2: Pseudo-masked language models for unified language model pre-training", *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Vienna, Austria, pp. 642–652.

Barrett, L.F. (2006), "Solving the emotion paradox: Categorization and the experience of emotion", *Personality and Social Psychology Review*, Sage Publications Sage CA: Los Angeles, CA, Vol. 10 No. 1, pp. 20–46, doi: 10.1207/s15327957pspr1001_2.

Borrajo-Millán, F., Alonso-Almeida, M.-M., Escat-Cortes, M. and Yi, L. (2021), "Sentiment analysis to measure quality and build sustainability in tourism destinations", *Sustainability*, MDPI, Vol. 13 No. 11, p. 6015, doi: 10.3390/su13116015.

Calderón-Fajardo, V., Anaya-Sánchez, R. and Molinillo, S. (2024), "Understanding destination brand experience through data mining and machine learning", *Journal of Destination Marketing and Management*, Vol. 31, p. 100862, doi: 10.1016/j.jdmm.2024.100862.

Cattell, R.B. (1940), "Sentiment or attitude? The core of a terminology problem in personality research", *Journal of Personality*, Vol. 9 No. 1, pp. 6–17, doi: 10.1111/j.1467-6494.1940.tb02192.x.

Chen, N., Hsu, C.H.C. and Li, X. (2021), "Resident sentiment toward a dominant tourist market: Scale development and validation", *Journal of Travel Research*, Vol. 60 No. 7, pp. 1408–1425, doi: 10.1177/0047287520947799.

Cheng, X., Chen, Y., Wang, P., Zhou, Y.X., Wei, X., Luo, W. and Duan, Q. (2024), "A novel ChatGPT-based multimodel framework for tourism review mining: a case study on China's five sacred mountains", *Journal of Hospitality and Tourism Technology*, Vol. 15 No. 4, pp. 592–609, doi: 10.1108/JHTT-06-2023-0170.

Costa, A., Guerreiro, J., Moro, S. and Henriques, R. (2019), "Unfolding the characteristics of incentivized online reviews", *Journal of Retailing and Consumer Services*, Elsevier Ltd, Vol. 47, pp. 272–281, doi: 10.1016/j.jretconser.2018.12.006.

Cui, Y., Che, W., Liu, T., Qin, B. and Yang, Z. (2021), "Pre-training with whole word masking for Chinese BERT", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, Vol. 29, pp. 3504–3514, doi: 10.1109/TASLP.2021.3124365.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1, ACL, Minneapolis, MN, pp. 4171–4186, doi: 10.48550/arXiv.1810.04805.

Fong, J.H., Murphy, T.D. and Pruitt, K.D. (2013), "Comparison of RefSeq protein-coding regions in human and vertebrate genomes", *BMC Genomics*, Vol. 14, p. 654, doi: 10.1186/1471-2164-14-654.

Fu, Y., Hao, J.X., Li, X. (Robert) and Hsu, C.H.C. (2019), "Predictive accuracy of sentiment analytics for tourism: A metalearning perspective on Chinese travel news", *Journal of Travel Research*, Vol. 58 No. 4, pp. 666–679, doi: 10.1177/0047287518772361.

Galesic, M. (2017), "We need more precise, quantitative models of sentiments", *Behavioral and Brain Sciences*, Vol. 40, p. e236, doi: 10.1017/S0140525X16000753.

Gaspar, R., Pedro, C., Panagiotopoulos, P. and Seibt, B. (2016), "Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events", *Computers in Human Behavior*, Vol. 56, pp. 179–191, doi: 10.1016/j.chb.2015.11.040.

Gervais, M.M. and Fessler, D.M.T. (2017), "On the deep structure of social affect: Attitudes, emotions, sentiments, and the case of 'contempt'", *Behavioral and Brain Sciences*, Vol. 40, p. e225, doi: 10.1017/S0140525X16000352.

Guo, X., Wang, Y., Tao, J. and Guan, H. (2024), "Identifying unique attributes of tourist

attractions: an analysis of online reviews", *Current Issues in Tourism*, Vol. 27 No. 3, pp. 479–497, doi: 10.1080/13683500.2023.2165904.

Hao, J.X., Fu, Y., Hsu, C., Li, X. and Chen, N. (2020), "Introducing news media sentiment analytics to residents' attitudes research", *Journal of Travel Research*, Vol. 59 No. 8, pp. 1353–1369, doi: 10.1177/0047287519884657.

Haslam, N. (2017), "A sentimental education: The place of sentiments in personality and social psychology", *Behavioral and Brain Sciences*, Vol. 40, p. e239, doi: 10.1017/S0140525X16000789.

Hsu, C.H.C., Li, X.R. and Chen, N. (2016), "Resident sentiment: Preliminary conceptualization and measurement", *2016 TTRA International Conference*.

Hsu, C.H.C., Tan, G. and Stantic, B. (2024), "A fine-tuned tourism-specific generative AI concept", *Annals of Tourism Research*, Elsevier Ltd, Vol. 104, p. 103723, doi: 10.1016/j.annals.2023.103723.

Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L. and Levy, O. (2020), "Spanbert: Improving pre-training by representing and predicting spans", *Transactions of the Association for Computational Linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, Vol. 8, pp. 64–77.

León, C.J., Suárez-Rojas, C., Cazorla-Artiles, J.M. and González Hernández, M.M. (2025), "Satisfaction and sustainability concerns in whale-watching tourism: A user-generated content model", *Tourism Management*, Vol. 106, p. 105019, doi: 10.1016/j.tourman.2024.105019.

Levis, M., Leonard Westgate, C., Gui, J., Watts, B. V. and Shiner, B. (2021), "Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models", *Psychological Medicine*, Vol. 51, pp. 1382–1391, doi: 10.1017/S0033291720000173.

Li, H., Yu, B.X.B., Li, G. and Gao, H. (2023), "Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews", *Tourism Management*, Elsevier Ltd, Vol. 96, p. 104707, doi: 10.1016/j.tourman.2022.104707.

Li, Z., Yuan, F. and Zhao, Z. (2025), "Robot restaurant experience and recommendation behaviour: based on text-mining and sentiment analysis from online reviews", *Current*

Issues in Tourism, Vol. 28 No. 3, pp. 461–475, doi: 10.1080/13683500.2024.2309140.

Liu, J. and Hu, S. (2023), "Text classification in tourism and hospitality – a deep learning perspective", *International Journal of Contemporary Hospitality Management*, Vol. 35 No. 12, pp. 4177–4190, doi: 10.1108/IJCHM-07-2022-0913.

Liu, J., Yu, Y., Mehraliyev, F., Hu, S. and Chen, J. (2022), "What affects the online ratings of restaurant consumers: a research perspective on text-mining big data analysis", *International Journal of Contemporary Hospitality Management*, Vol. 34 No. 10, pp. 3607–3633, doi: 10.1108/IJCHM-06-2021-0749.

Liu, Y., Huang, K., Bao, J. and Chen, K. (2019), "Listen to the voices from home: An analysis of Chinese tourists' sentiments regarding Australian destinations", *Tourism Management*, Elsevier, Vol. 71, pp. 337–347, doi: 10.1016/j.tourman.2018.10.004.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., *et al.* (2019), "RoBERTa: a robustly optimized BERT pretraining approach", *Computation and Language*, available at: https://arxiv.org/abs/1907.11692 (accessed 21 August 2025).

Luo, Y. and Xu, X. (2021), "Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic", *International Journal of Hospitality Management*, Elsevier Ltd, Vol. 94, p. 102849, doi: 10.1016/j.ijhm.2020.102849.

Mahmoud, A.B., Fuxman, L., Asaad, Y. and Solakis, K. (2025), "Exploring new realms or losing touch? Assessing public beliefs about tourism in the metaverse–a big-data approach", *International Journal of Contemporary Hospitality Management*, Vol. 37 No. 4, pp. 1384–1420, doi: 10.1108/IJCHM-09-2023-1515.

Mariani, M. and Borghi, M. (2021), "Are environmental-related online reviews more helpful? A big data analytics approach", *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 6, pp. 2065–2090, doi: 10.1108/IJCHM-06-2020-0548.

Mehraliyev, F., Chan, I.C.C. and Kirilenko, A.P. (2022), "Sentiment analysis in hospitality and tourism: a thematic and methodological review", *International Journal of Contemporary Hospitality Management*, Vol. 34 No. 1, pp. 46–77, doi: 10.1108/IJCHM-02-2021-0132.

Mehraliyev, F., Kirilenko, A.P. and Choi, Y. (2020), "From measurement scale to sentiment scale: Examining the effect of sensory experiences on online review rating behavior",

Tourism Management, Vol. 79, p. 104096, doi: 10.1016/j.tourman.2020.104096.

Munezero, M., Montero, C.S., Sutinen, E. and Pajunen, J. (2014), "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text", *IEEE Transactions on Affective Computing*, IEEE, Vol. 5 No. 2, pp. 101–111, doi: 10.1109/TAFFC.2014.2317187.

Naar, H. (2013), *A Defence of Sentiments: Emotions, Dispositions, and Character*, The University of Manchester (United Kingdom).

Oscar, N., Fox, P.A., Croucher, R., Wernick, R., Keune, J. and Hooker, K. (2017), "Machine learning, sentiment analysis, and tweets: An examination of Alzheimer's disease stigma on Twitter", *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, Vol. 72 No. 5, pp. 742–751, doi: 10.1093/geronb/gbx014.

Pascanu, R., Mikolov, T. and Bengio, Y. (2013), "On the difficulty of training recurrent neural networks", *Proceedings of the 30th International Conference on Machine Learning*, Pmlr, Atlanta, Georgia, pp. 1310–1318.

Peng, H. gang, Zhang, H. yu and Wang, J. qiang. (2018), "Cloud decision support model for selecting hotels on TripAdvisor.com with probabilistic linguistic information", *International Journal of Hospitality Management*, Elsevier, Vol. 68, pp. 124–138, doi: 10.1016/j.ijhm.2017.10.001.

Qiao, T., Shan, W., Zhang, M. and Wei, Z. (2022), "More than words: Understanding how valence and content affect review value", *International Journal of Hospitality Management*, Elsevier Ltd, Vol. 105, p. 103274, doi: 10.1016/j.ijhm.2022.103274.

Rita, P., Vong, C., Pinheiro, F. and Mimoso, J. (2023), "A sentiment analysis of Michelin-starred restaurants", *European Journal of Management and Business Economics*, Emerald Publishing Limited, Vol. 32 No. 3, pp. 276–295.

Russell, J.A. (1980), "A circumplex model of affect.", *Journal of Personality and Social Psychology*, American Psychological Association, Vol. 39 No. 6, p. 1161.

Saleh, M.I. (2025), "Generative artificial intelligence in hospitality and tourism: Future capabilities, AI prompts and real-world applications", *Journal of Hospitality Marketing & Management*, Routledge, pp. 1–32, doi: 10.1080/19368623.2025.2458603.

Seyfi, S., Kim, M.J., Nazifi, A., Murdy, S. and Vo-Thanh, T. (2025), "Understanding tourist

barriers and personality influences in embracing generative AI for travel planning and decision-making", *International Journal of Hospitality Management*, Elsevier Ltd, Vol. 126 No. May 2024, p. 104105, doi: 10.1016/j.ijhm.2025.104105.

Shand, A.F. (1922), "The relations of complex and sentiment. III", *British Journal of Psychology. General Section*, Vol. 13 No. 2, pp. 123–129, doi: 10.1111/j.2044-8295.1922.tb00088.x.

Shenzhen Shukuo Information Technology Co. Ltd. (2013), "Octopus", available at: https://octopus.com/ (accessed 5 July 2025).

Shin, S. and Nicolau, J.L. (2022), "Identifying attributes of wineries that increase visitor satisfaction and dissatisfaction: Applying an aspect extraction approach to online reviews", *Tourism Management*, Elsevier Ltd, Vol. 91, p. 104528, doi: 10.1016/j.tourman.2022.104528.

Statista. (2024), "Main barriers that prevented travel companies from implementing generative artificial intelligence (AI) worldwide as of 3rd quarter 2024", available at: https://www.statista.com/statistics/1500104/main-barriers-generative-ai-adoption-travel-companies-worldwide/ (accessed 3 February 2025).

Tetzlaff, L., Rulle, K., Szepannek, G. and Gronau, W. (2019), "A customer feedback sentiment dictionary: Towards automatic assessment of online reviews", *European Journal of Tourism Research*, Varna University of Management, Vol. 23, pp. 28–39.

Vogklis, K. and Gkritzali, A. (2024), "The risk of local crises for destination image", *Annals of Tourism Research*, Elsevier Ltd, Vol. 107, p. 103797, doi: 10.1016/j.annals.2024.103797.

Wei, Z., Zhang, M. and Ming, Y. (2023), "Understanding the effect of tourists' attribute-level experiences on satisfaction–a cross-cultural study leveraging deep learning", *Current Issues in Tourism*, Vol. 26 No. 1, pp. 105–121, doi: 10.1080/13683500.2022.2030682.

Wu, D.C., Zhong, S., Song, H. and Wu, J. (2024), "Do topic and sentiment matter? Predictive power of online reviews for hotel demand forecasting", *International Journal of Hospitality Management*, Vol. 120, p. 103750, doi: 10.1016/j.ijhm.2024.103750.

Wu, J., Chen, J., Yang, T. and Zhao, N. (2024), "How to stay competitive: An innovative concept to assess the business competitiveness using online restaurant reviews", *International Journal of Hospitality Management*, Elsevier Ltd, Vol. 122, p. 103836, doi:

10.1016/j.ijhm.2024.103836.

Wu, J. and Yang, T. (2023), "Service attributes for sustainable rural tourism from online comments : Tourist satisfaction perspective", *Journal of Destination Marketing & Management*, Elsevier Ltd, Vol. 30 No. 2, p. 100822, doi: 10.1016/j.jdmm.2023.100822.

Wu, J., Yang, T., Zhou, Z. and Zhao, N. (2023), "Consumers' affective needs matter: Open innovation through mining luxury hotels' online reviews", *International Journal of Hospitality Management*, Elsevier Ltd, Vol. 114, p. 103556, doi: 10.1016/j.ijhm.2023.103556.

Wu, J., Zhang, C.J., Huang, G.I., Yang, T. and Hao, F. (2025), "Taste, trend, and turmoil: Tracking the life cycle of internet-famous restaurants through customer satisfaction", *International Journal of Hospitality Management*, Elsevier Ltd, Vol. 126, p. 104071, doi: 10.1016/j.ijhm.2024.104071.

Wu, J. and Zhao, N. (2023), "What consumer complaints should hoteliers prioritize? Analysis of online reviews under different market segments", *Journal of Hospitality Marketing and Management*, Routledge, Vol. 32 No. 1, pp. 1–28, doi: 10.1080/19368623.2022.2119187.

Wu, J., Zhao, N. and Yang, T. (2024), "Wisdom of crowds: SWOT analysis based on hybrid text mining methods using online reviews", *Journal of Business Research*, Elsevier Inc., Vol. 171, p. 114378, doi: 10.1016/j.jbusres.2023.114378.

Yang, T. and Hsu, C.H.C. (2025), "Integrating community ambivalence into resident sentiment research", *Journal of Travel Research*, doi: 10.1177/00472875251366433.

Yang, T., Wu, J. and Zhang, J. (2024), "Knowing how satisfied/dissatisfied is far from enough: a comprehensive customer satisfaction analysis framework based on hybrid text mining techniques", *International Journal of Contemporary Hospitality Management*, Vol. 36 No. 3, pp. 873–892, doi: 10.1108/IJCHM-10-2022-1319.

Yang, T., Zhang, C.J., Wu, J. and Zhang, J. (2025), "The hidden gems in online reviews : unraveling how the expressions of affective needs impact review usefulness", *Journal of Hospitality Marketing & Management*, Routledge, Vol. 34 No. 2, pp. 234–256, doi: 10.1080/19368623.2024.2413189.

Yin, X. and Jung, T. (2024), "Analysing the causes of tourists' emotional experience related to tourist attractions from a binary emotions perspective utilising machine learning models",

Asia Pacific Journal of Tourism Research, Vol. 29 No. 6, pp. 699–718, doi: 10.1080/10941665.2024.2343077.

Zainal, N.H., Eckhardt, R., Rackoff, G.N., Fitzsimmons-Craft, E.E., Rojas-Ashe, E., Barr Taylor, C., Funk, B., *et al.* (2025), "Capitalizing on natural language processing (NLP) to automate the evaluation of coach implementation fidelity in guided digital cognitive-behavioral therapy (GdCBT)", *Psychological Medicine*, Vol. 55 No. e106, pp. 1–13, doi: 10.1017/S0033291725000340.

Appendices

Appendix A: Python codes

```
import pandas as pd
import torch
from torch.utils.data import Dataset, DataLoader
from transformers import AutoTokenizer, AutoModel, AdamW, get_linear_schedule_with_warmup
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import os

# Set device
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

debug = False

# Read data
data_path = 'test/Sentiment annotation-final version-test.xlsx' if debug else "data/Sentiment annotation-final
version.xlsx"
data = pd.read_excel(data_path)

# Extract text and labels
texts = data['Text'].tolist()
labels = data['Final label'].tolist()

# divide the dataset
train_texts, val_texts, train_labels, val_labels = train_test_split(texts, labels, test_size=0.1, random_state=42)

# Initialize tokenizer and model
model_name = 'hfl/chinese-roberta-wwm-ext'
tokenizer = AutoTokenizer.from_pretrained(model_name)

# Load original model
model = AutoModel.from_pretrained(model_name)
model = model.to(device)

# Customize the dataset
class SentimentDataset(Dataset):
    def __init__(self, texts, labels, tokenizer, max_len):
        self.texts = texts
        self.labels = labels
        self.tokenizer = tokenizer
```

```
            self.max_len = max_len

        def __len__(self):
            return len(self.texts)

        def __getitem__(self, item):
            text = str(self.texts[item])
            label = self.labels[item]

            encoding = self.tokenizer.encode_plus(
                text,
                add_special_tokens=True,
                max_length=self.max_len,
                return_token_type_ids=False,
                padding='max_length',
                truncation=True,
                return_attention_mask=True,
                return_tensors='pt',
            )

            return {
                'input_ids': encoding['input_ids'].flatten(),
                'attention_mask': encoding['attention_mask'].flatten(),
                'label': torch.tensor(label, dtype=torch.float)
            }

# Set super parameters
MAX_LEN = 64
BATCH_SIZE = 64 if debug else 256
EPOCHS = 100
LEARNING_RATE = 2e-5
WARMUP_STEPS = 1000
PATIENCE = 3 if debug else 10    # Early stopping patience

# Create dataset and DataLoader
train_dataset = SentimentDataset(train_texts, train_labels, tokenizer, MAX_LEN)
val_dataset = SentimentDataset(val_texts, val_labels, tokenizer, MAX_LEN)

train_loader = DataLoader(train_dataset, batch_size=BATCH_SIZE, shuffle=True)
val_loader = DataLoader(val_dataset, batch_size=BATCH_SIZE, shuffle=False)

# Define the class of model
class RobertaRegressor(torch.nn.Module):
    def __init__(self, roberta_model):
```

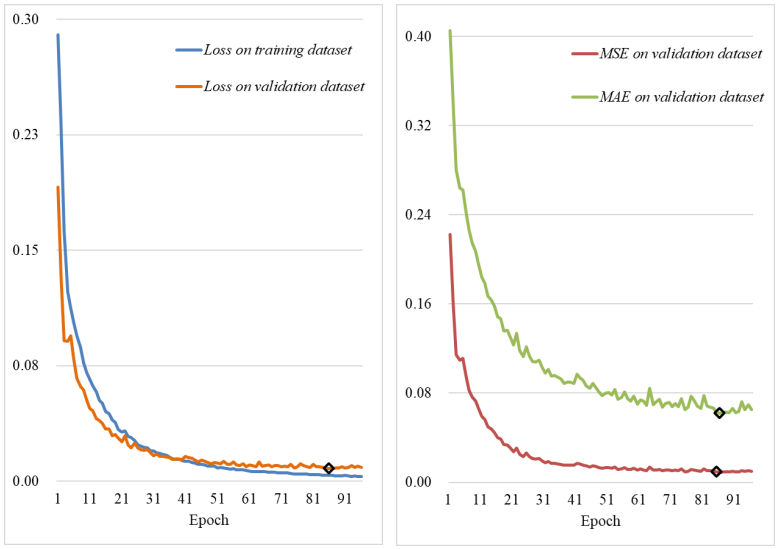
```
            super(RobertaRegressor, self).__init__()
            self.roberta = roberta_model
            self.regressor = torch.nn.Linear(roberta_model.config.hidden_size, 1)

        def forward(self, input_ids, attention_mask):
            outputs = self.roberta(input_ids=input_ids, attention_mask=attention_mask)
            pooled_output = outputs.last_hidden_state[:, 0, :]  # extract [CLS]
            output = self.regressor(pooled_output)
            return output.squeeze()

# Initialize the model
model = RobertaRegressor(model).to(device)

# Define loss function and optimizer
loss_fn = torch.nn.MSELoss()
optimizer = AdamW(model.parameters(), lr=LEARNING_RATE)
total_steps = len(train_loader) * EPOCHS
scheduler = get_linear_schedule_with_warmup(optimizer, num_warmup_steps=WARMUP_STEPS,
num_training_steps=total_steps)

# Record the optimal model
best_mse = float('inf')
best_epoch = 0
best_loss = float('inf')
best_metrics = None
early_stopping_counter = 0  # Early stopping

# Training and validation
results = {'Epoch': [], 'Train Loss': [], 'Train MSE': [], 'Train MAE': [], 'Train R2': [],
            'Val Loss': [], 'Val MSE': [], 'Val MAE': []}

for epoch in range(EPOCHS):
    # training
    model.train()
    train_loss = 0
    train_predictions, train_true_labels = [], []
    for batch in train_loader:
        input_ids = batch['input_ids'].to(device)
        attention_mask = batch['attention_mask'].to(device)
        labels = batch['label'].to(device)

        optimizer.zero_grad()
        outputs = model(input_ids, attention_mask)
        loss = loss_fn(outputs, labels)
```

```
            loss.backward()
            optimizer.step()
            scheduler.step()

            train_loss += loss.item()
            train_predictions.extend(outputs.detach().cpu().numpy())
            train_true_labels.extend(labels.cpu().numpy())

        avg_train_loss = train_loss / len(train_loader)
        train_mse = mean_squared_error(train_true_labels, train_predictions)
        train_mae = mean_absolute_error(train_true_labels, train_predictions)
        train_r2 = r2_score(train_true_labels, train_predictions)

        # validation
        model.eval()
        val_loss = 0
        val_predictions, val_true_labels = [], []
        with torch.no_grad():
            for batch in val_loader:
                input_ids = batch['input_ids'].to(device)
                attention_mask = batch['attention_mask'].to(device)
                labels = batch['label'].to(device)

                outputs = model(input_ids, attention_mask)
                loss = loss_fn(outputs, labels)

                val_loss += loss.item()
                val_predictions.extend(outputs.cpu().numpy())
                val_true_labels.extend(labels.cpu().numpy())

        avg_val_loss = val_loss / len(val_loader)
        val_mse = mean_squared_error(val_true_labels, val_predictions)
        val_mae = mean_absolute_error(val_true_labels, val_predictions)

        print(f'Epoch: {epoch+1}/{EPOCHS}')
        print(f'Training Loss: {avg_train_loss:.4f}, Training MSE: {train_mse:.4f}, Training MAE: {train_mae:.4f}')
        print(f'Validation Loss: {avg_val_loss:.4f}, Validation MSE: {val_mse:.4f}, Validation MAE: {val_mae:.4f}')

        # Record the results
        results['Epoch'].append(epoch + 1)
        results['Train Loss'].append(avg_train_loss)
        results['Train MSE'].append(train_mse)
```

```
        results['Train MAE'].append(train_mae)
        results['Train R2'].append(train_r2)
        results['Val Loss'].append(avg_val_loss)
        results['Val MSE'].append(val_mse)
        results['Val MAE'].append(val_mae)

        # Early stopping mechanism
        if avg_val_loss < best_loss:
            best_mse = val_mse
            best_epoch = epoch + 1
            best_loss = avg_val_loss
            best_metrics = {
                'Loss': best_loss,
                'MSE': val_mse,
                'MAE': val_mae
            }
            torch.save(model.state_dict(), 'test/roberta-best_model-test_loss_' + str(best_loss) + '.pth' if debug
else 'model/roberta-best_model_loss_' + str(best_loss) + '.pth')
            early_stopping_counter = 0    # Reset counter
        else:
            early_stopping_counter += 1
            if early_stopping_counter >= PATIENCE:
                print(f'Early stopping triggered after {PATIENCE} epochs without improvement.')
                break

# Output the optimal results
print(f'Best MSE: {best_mse:.4f} at epoch {best_epoch} with loss {best_loss:.4f}')
print(f'Best metrics: {best_metrics}')

# Save
results_df = pd.DataFrame(results)
if not os.path.exists('result/training'):
    os.makedirs('result/training')
results_df.to_excel('test/training/roberta-test_loss_' + str(best_loss) + '.xlsx' if debug
                    else 'result/training/roberta_loss_' + str(best_loss) + '.xlsx', index=False)
```

Appendix B: The training process

Based on the proposed metrics, model training procedures were carried out. Referring to extant studies (e.g., Okafor *et al.*, 2017; Pilania *et al.*, 2015), the annotated dataset was randomly divided into two sub-datasets of 0.9:0.1, with the former serving as the training dataset and the latter as the validation dataset. Figure B1 presents the values of different metrics throughout the training process. Lower values are the optimal values of *Loss, MSE*, and *MAE*. Based on Equation (6), Figure B1 (a) shows *Loss* values on the training dataset. The two lines in Figure B1 (a) display different trends, with *Loss* value on the validation dataset first decreasing as the epoch increased, but then began to increase from Epoch 86. Thus, the optimal value of epoch is on Epoch 86. In contrast, *Loss* value on the training dataset continued to decrease as the epoch increased. This reflects the phenomenon that as the model continued to be trained, features of the training dataset were gradually learned by the model, and therefore, the *Loss* function of the model on the training dataset kept decreasing to the point of "overfitting". Figure B1 (b) illustrates the changes of *MSE* and *MAE*. Aligning with the change of *Loss* in Figure B1 (a), *MSE* and *MAE* also perform best on Epoch 86.



(a) *Loss* on training and validation dataset (b) *MSE* and *MAE* on validation dataset

Figure B1 Visualization of the training process

Note: ◊ *in the Figure indicates the optimal value of each metric*

After training the model, we obtained the optimal parameters of the model structure. Table B1 presents key parameters, such as the max length of input sequence (i.e., 64 characters), batch size when training the model (i.e., 256 pieces of data), and the learning rate (0.00002). Moreover, the experimental environment can be seen in Table B1, including the deep learning framework used (i.e., PyTorch) and the version of the basic model (i.e., RoBERTa/chinese-roberta-wwm-ext).

Table B1. Experimental environment and configuration

Experimental environment	Parameter
Programming language	Python 3.9.19
Deep learning framework	PyTorch
RoBERTa	chinese-roberta-wwm-ext
Local development environment	PyCharm Community Edition 2024.1
Memory	64G
GPU	NVIDIA GeForce RTX 3060 Ti
Max_len	64
Batch_size	256
Warmup steps	1,000
Learning rate	0.00002
Optimizer	Torch.optim.AdamW

1

2

Appendix C: Validation on an English dataset

To further validate the proposed model, we collected all tourists' online reviews of a famous tourist attraction in Hong Kong (The Peak) from three online platforms, including TripAdvisor.com, Ctrip.com, and Booking.com. Since the dataset in the main study is about a hotel brand, we purposely collected data in the context of tourism to enhance the replicability of our proposed method. Additionally, we collected data from various OTAs, aiming at further strengthening the generalizability of this research. Overall, for validation, we obtained 22,440 pieces of English online reviews.

Aligning with the rules of sentence segmenting in the main study, we segmented the online reviews in the English dataset into sentences. Furthermore, we randomly selected 1,000 data samples from the sentences. Two tourism researchers were recruited to annotate the English sentences, using the same rules explained in the manuscript (Pearson correlation coefficients=0.7894). The two researchers' annotation results are considered to construct the standard annotated dataset. Then, we adopted NLLB-200 (No Language Left Behind), which is an advanced large language model developed by Facebook (https://huggingface.co/facebook/nllb-200-distilled-600M), to translate the English data into Chinese. Last, we re-ran the models used for comparison, as well as our proposed model. Table C1 presents the comparison results, which show that our proposed model achieves the best performance regarding metric *MSE* and slightly higher *MAE* than ChatGPT-4o.

Table C1. Comparison of results from an English dataset

Type	**Model**	**MSE**	**MAE**	**Cost**
Open-source packages	SnowNLP	0.13	0.29	Free
	Paddle	0.08	0.23	Free
Deep learning models	CNN	0.06	0.20	Free
	LSTM	0.06	0.21	Free
Pre-trained language models	ERNIE + regressor layer	0.06	0.19	Free
	XLNet + regressor layer	0.06	0.19	Free
	BERT + regressor layer	0.05	0.17	Free
Generative AI tools	Qwen-long	0.04	0.14	$1 / 1M tokens
	ChatGPT-4o	0.04	**0.13**	$2.5 / 1M tokens
Proposed model	**RoBERTa-CSS**	**0.04**	0.14	Free

Appendix D: Validation on the dataset with noise

Since user-generated reviews often contain noise such as typographical errors, ambiguous statements, or incomplete sentences, which can challenge the model's predictive accuracy, we further tested the proposed model on a dataset with noise. Specifically, we selected 100 sentences with the above noise from the original labelled dataset as the test dataset. Figure D1 shows the results regarding *MSE* and *MAE*. From the figure, the optimal epoch number is 84, and the best value of *MSE* and *MAE* is 0.01 and 0.08, respectively. Compared to the results obtained with pre-processed data as in Table 1 (i.e., *MSE* equals 0.01 and *MAE* equals 0.06), the proposed model performs well on noisy data.

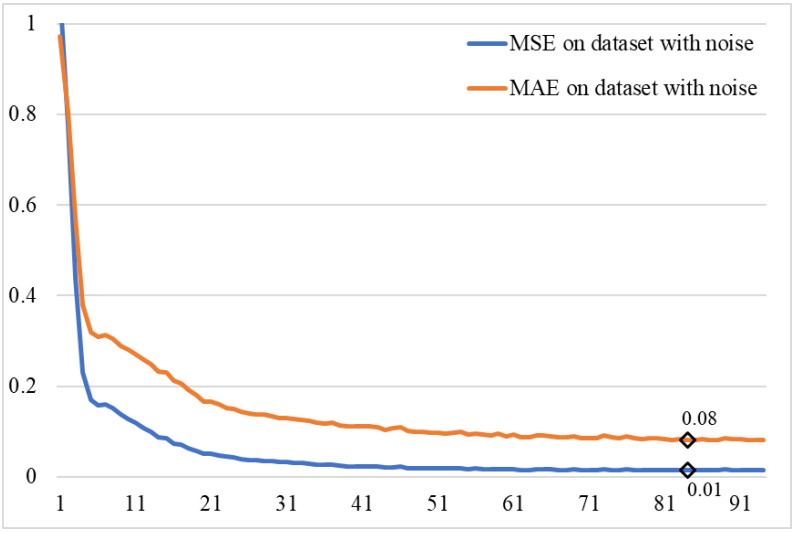


Figure D1 *MSE* and *MAE* on dataset with noise

Appendix E: Group-level longitudinal analysis

Time series analysis of tourist sentiment can also benefit from our proposed approach to estimate tourists' continuous sentiment scores. Based on the calculated results of BERT and the proposed model using the full dataset, we can illustrate dynamic changes in tourist sentiment, as shown in Figure E1, where the line chart is based on RoBERTa-CSS, and the bar chart is based on the BERT model.

Taking the period of 2019-2021 as an example (i.e., the part enclosed by the red rectangle), negative labels (i.e., brown bars) are decreasing, and positive labels (i.e., green bars) are increasing, which seems to imply that customers were feeling more positive over time. However, the trend of continuous sentiment scores (i.e., the black line) indicates that customers reported a lower level of positive sentiment (> 0.5), which cannot be observed by extant sentiment classification models producing categorical results. Such a comparison demonstrated the necessity of developing an accurate approach to measure tourist sentiment using continuous scoring.

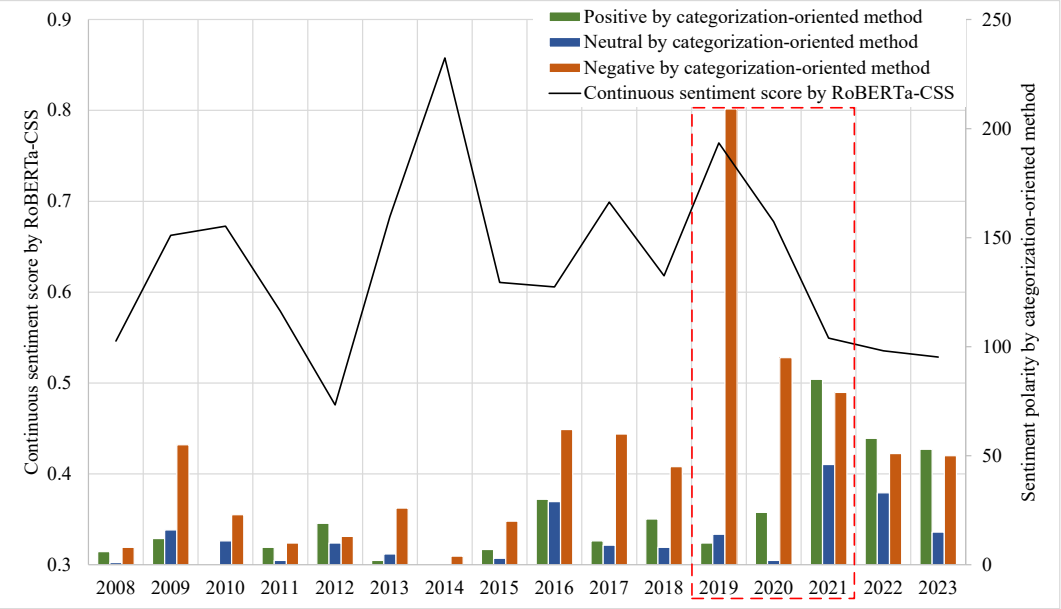


Figure E1 Comparison of time series (by RoBERTa-CSS vs. the categorization-oriented method)

Note: Line chart is based on proposed method; Bar chart is based on the categorization-oriented method

These analyses show that RoBERTa-CSS could enable the marketing and communications (MarCom) department to identify more appropriate timing for relevant strategies. Figure E1 shows that customer continuous sentiments decreased sharply in 2010-2012, 2014-2015, and 2019-2021; thus, proactive strategies could be implemented at these timepoints. If the continuous sentiment value is low, the MarCom team should be more proactive in communicating with consumers and offering targeted compensation for service failures based on user profiles, such as public apologies or room upgrades. In other words, the MarCom team can adopt proactive marketing and reputation management activities based on the changing trends of customer continuous sentiments.

References

Okafor, E., Pawara, P., Karaaba, F., Surinta, O., Codreanu, V., Schomaker, L. and Wiering, M. (2017), "Comparative study between deep learning and bag of visual words for wild-animal recognition", *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, IEEE, pp. 1–8, doi: 10.1109/SSCI.2016.7850111.

Pilania, G., Gubernatis, J.E. and Lookman, T. (2015), "Structure classification and melting temperature prediction in octet AB solids via machine learning", *Physical Review B - Condensed Matter and Materials Physics*, Vol. 91 No. 21, pp. 1–13, doi: 10.1103/PhysRevB.91.214302.