Asia Pacific Academy of Science Pte. Ltd.

Smart Tourism 2024, 5(2), 2700.
https://doi.org/10.54517/st.v5i2.2700

Article

Construction of the tourist sentiment dictionary for hotels to mining tourist demands: Based on Macao's hotel reviews from Agoda

Linyu Wang[1], Xiaohan Zhu[2], Hongbin Zhang[3], Chenhe Zhang[4], Jiajun Xu[5,*], Zhaochen Zhang[6]

[1] School of Drama and Film, Jilin University of Arts, 130021 Changchun, China
[2] School of Humanities, Jiangnan University, 214122 Wuxi, China
[3] School of Management, Zhejiang University, 310058 Hangzhou, China
[4] School of Arts and Design, Yanshan University, 066004 Qinhuangdao, China
[5] Institute for Advanced Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia
[6] School of Geosciences, China University of Petroleum, 266580 Qingdao, China
*** Corresponding author:** Jiajun Xu; jiajunxu2000@gmail.com

CITATION

Wang L, Zhu X, Zhang H, et al. Construction of the tourist sentiment dictionary for hotels to mining tourist demands: Based on Macao's hotel reviews from Agoda. Smart Tourism. 2024; 5(2): 2700. https://doi.org/10.54517/st.v5i2.2700

ARTICLE INFO

Received: 26 April 2024
Accepted: 27 June 2024
Available online: 6 July 2024

Abstract: Tourist hotels (or tourist accommodations) are located near tourist attractions, primarily serving tourists. In recent years, with the gradual improvement of people's living standards around the globe, tourists' demands and standards for tourist hotel construction have been rising accordingly. In the context of technologization and informatization, various hotel booking platforms (Agoda, Booking, Trip, etc.) cover a large amount of review data in evaluating systems to reflect tourists' demands. Meanwhile, identifying demand-oriented reviews and extracting core consumer demands from them is crucial for optimizing hotel services and enhancing tourist satisfaction. Therefore, this study explores the demands of tourists in tourist hotels from the perspective of text sentiment analysis and takes Macao, a famous tourist destination, as an example, based on reviews of tourist hotels on the Agoda site platform. Specifics are as follows: (1) Based on pointwise mutual information (PMI) and information entropy (IE), it realizes the identification of sentiment words in the field of tourist hotels and constructs a sentiment dictionary to address the problem of poor relevance between word segmentation results; (2) It summarizes the five types of reviews containing tourist demands (positive, negative, suggestion, demand, and comparison) and their characteristics to solve the ambiguity of texts and further accurately reveal the main demands of tourists; (3) It classifies tourist demands and group similar tourist demands into the same categories to address the problem of multiple expressions for the same demand. The present study provides empirical experiences from Macao's hotels and contributes to the literature on text sentiment analysis in tourist hotels. Furthermore, the study results could enhance the mining accuracy and provide a detailed summarization of consumer demands and directions for the sustainable optimization improvement of tourist services.

Keywords: tourist hotel; text sentiment analysis; sentiment dictionary; demand mining; agoda platform; Macao of China

1. Introduction

The tourist hotel, also known as tourist accommodation, refers to commercial establishments that provide accommodations, dining, and other services for tourists [1,2]. These hotels are typically located near tourist attractions or transportation hubs and offer various types of accommodation options, facilities, and services to meet the diverse demands of tourists, helping them enjoy the pleasures of tourism and explore their destinations [3]. With the approaching end of the COVID-19 pandemic, governments worldwide are gradually relaxing prevention and control measures for

the epidemic, and the travel and tourism sector demonstrates robust indications of resurgence [4,5]. Global tourists exceeded 900 million in 2022, a two-fold increase compared to 2021 [6]. In 2022, international tourism revenue returned to US$ 1 trillion, a real increase of 50% compared with 2021 [6]. The global tourism development in the post-epidemic era is showing a positive trend, and the standard of tourism-related infrastructure and supporting facilities is continuously being raised by the growing consumer demand for tourism [7]. It is projected that the revenue of the tourist hotel market will reach US$446.50 billion by 2024 worldwide [8]. Looking forward, an annual growth rate of 3.32% (CAGR 2024–2028) is expected, resulting in a projected market volume of US$508.90 billion by 2028 [8].

When choosing a tourist hotel, tourists consider various factors, such as the environment [9], price [10], transportation [11], services and facilities provided by the hotel [12], and service attitude [13]. A more comfortable environment could put visitors in a better mood, affecting the subjective feelings of the hotel occupants and their evaluation of the environmental experience [14,15]. According to Environmental, Social, and Governance (ESG) theory, hotels that value guests' feedback and meet guest demands could positively influence guests' subjective choices, brand preferences, emotional trust, and post-use evaluations [16], thereby contributing to the hotel's sustainable development [17]. Nowadays, driven by machine learning, information models, and big data from the internet, there is an increasing amount of semantically rich text review data. The complexity of text data is leading people to shift from traditional manual visual screening to computer-based algorithmic model construction and intelligent evaluation, aiming to maximize the efficiency and reliability of data processing workflows [18–20]. Online hotel booking platforms contain vast amounts of reviews from tourists [21], including information about sentiments [22], demands [23], and improvement suggestions [24]. If improvement suggestions and demands from reviews could be accurately extracted, tourist hotels could optimize their services and enhance tourist satisfaction based on tourist feedback and demands. This, in turn, contributes to the sustainability of the tourism industry [25].

In the above context, this paper extracts tourist demands from the reviews of tourist hotels on online hotel booking platforms and provides suggestions for the improvement and optimization of tourist hotels. After a comprehensive literature review, this paper highlights the issues as well as challenges that have not yet been addressed by text sentiment analysis in the field of consumer demand feedback in tourist hotels. The lack of a sentiment dictionary suitable for the field of tourist hotels resulted in poor relevance between word segmentation results and the field of tourist hotels [26], the ambiguity of texts made it difficult to identify demands [27], and the same demand may have multiple expressions [28]. To address the problems mentioned above, this paper conducts the following research: (1) designing a sentiment word recognition algorithm based on Pointwise Mutual Information (PMI) [29] and Information Entropy (IE) [30] to identify sentiment words in the field of tourist hotels and construct a sentiment dictionary, ensuring that the word segmentation results align with this field; (2) summarizing the types of reviews containing tourist demands and their characteristics to solve the problem of ambiguity of texts; (3) accurately identifying tourist demands and group similar tourist demands into the same categories

to address the problem of multiple expressions for the same demand, providing references for research on text sentiment analysis in this domain [31]. This study proposes the innovatively optimal application of the text sentiment analysis method in tourist hotel services, and the study findings could contribute to revealing the important factors affecting consumer satisfaction in tourist hotels, as well as providing directions for the smart and sustainable optimization improvement of tourist services.

2. Materials and methods

2.1. Description of the study area

Macao, also known as Macau, is a small region along the southern coast of China, covering approximately 29.5 square kilometers [32]. It is located in the Pearl River Delta region, bordered by Zhuhai of Guangdong Province to the west and north and facing the South China Sea to the east and south (**Figure 1**). In the 16th century, Portuguese merchants primarily settled in Macao, which was a Portuguese colony from 1887 to 1999 [32]. On December 20, 1999, sovereignty over Macao was officially transferred back to China, becoming one of China's two Special Administrative Regions [32].

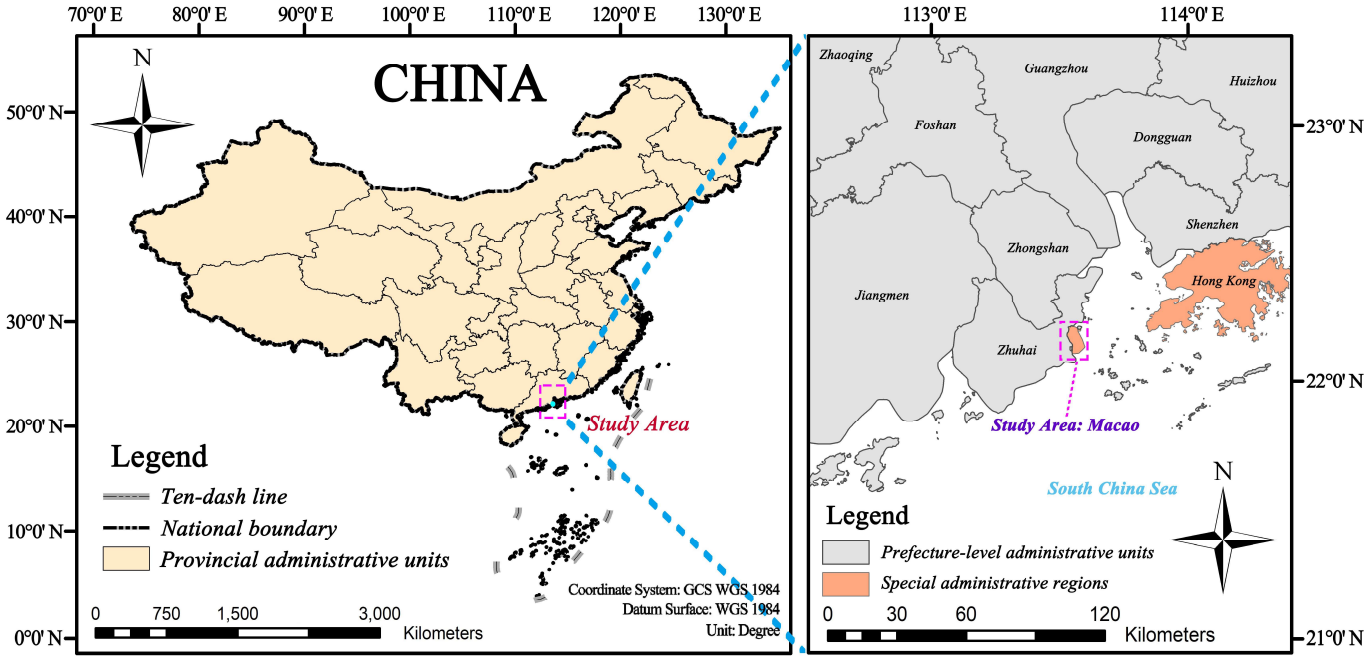

Figure 1. Administrative scope and geographical location of Macao [Note: The map is based on the Chinese drawing review No: GS (2023) 2762].

Due to its economy being mainly driven by gambling, tourism, and speculative investments, the economic capital of Macao has taken on a mysterious form to generate huge profits, attracting millions of visitors, especially mainland Chinese tourists from the border and foreign tourists who come out of curiosity [33]. Among them, gambling and historical tradition are two key engines driving the transformation of the city of Macao and the surge in Gross Domestic Product (GDP) [33,34]. Their unexpected combination has made this city an attractive and renowned tourist destination, especially after the Historic Centre of Macao was listed as a UNESCO

(United Nations Educational, Scientific and Cultural Organization) World Heritage Site in 2005 [33,34].

2.2. Data collection

This study selected hotels and all the tourist reviews of Macao hotels as of April 20, 2024, to construct the dataset. The dataset consists of two dimensions: hotel and reviews. The hotel dimension includes three attributes: hotel order, hotel name, and hotel score. The review dimension includes three attributes: review title, content, and tourist score. The explanations of these attributes are shown in **Table 1**.

Table 1. The attributes of data and their meanings.

Dimension	Attributes	Meanings
Hotel	Hotel Order	The unique order of the hotel.
	Hotel Name	The name of the hotel.
	Hotel Score	The average scores given by tourists who have stayed in the hotel, with scores ranging from 0 to 10. The scores include five dimensions: Cleanliness, Location, Value for money, Facilities, and Service.
Review	Review Title	The title of the tourists' review of their stay at the hotel (which usually contains brief information).
	Review Content	The content of the tourists' reviews of their stay at the hotel (which usually contains detailed information).
	Tourist Score	The scores given by the tourists for their stay ranged from 0 to 10, covering five dimensions: Cleanliness, Location, Value for money, Facilities, and Service.

Excluding hotels without English reviews, the dataset contains data from 86 hotels in Macao. The reason for only selecting English-language reviews is twofold: firstly, to ensure uniformity and comparability in text data processing; secondly, English reviews constitute the majority, thus providing a sufficient amount of text to guarantee the accuracy of sentiment word extraction [26]. For the study design of this paper, two datasets need to be constructed. (1) Dataset 1: English reviews for sentiment word identification and sentiment dictionary construction in the tourist hotel domain. (2) Dataset 2: Positive and negative English reviews for mining tourist demands. The following sections introduce these two datasets.

(1) Dataset 1

There are differences in grammar structures between different languages, which could have an impact on mining tourist demands based on text sentiment analysis methods in this study [35]. Therefore, this study chooses to use the most prevalent English reviews for text sentiment analysis. The statistical results of the ratio of English reviews for hotels in Dataset 1 are shown in **Figure 2**.

Dataset 1 contains 87,975 reviews. Based on **Figure 2**, it can be seen that the ratio of English reviews in most hotels is above 30%. Therefore, selecting English reviews to extract sentiment words and build a sentiment dictionary maximizes the amount of data available in this study, thus making sentiment word extraction more accurate [36]. Dataset 1 contains 87,975 reviews.

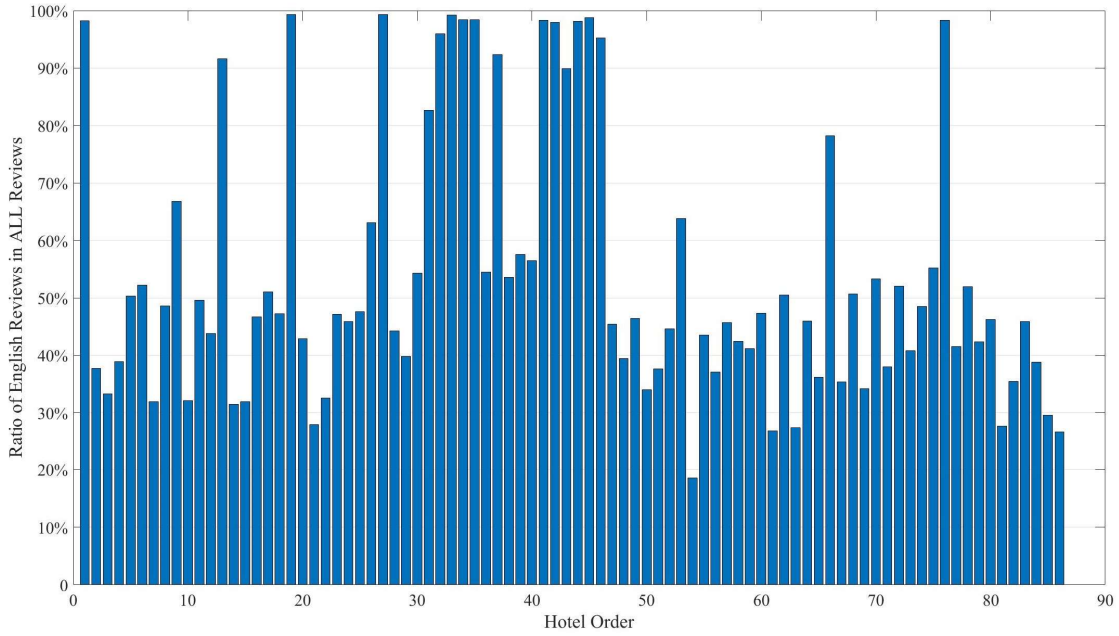

Figure 2. The ratio of English reviews in Dataset 1.

(2) Dataset 2

Based on the distribution of scores in the dataset and the general impression people have of scoring, reviews with scores exceeding 9 are defined as positive reviews, reviews with scores between 8 and 9 are defined as average reviews, and reviews with scores below 8 are defined as negative reviews. Since text with strong sentiment polarity, like positive and negative reviews, often contain people's opinions, viewpoints, and demand information [37], this study chooses positive and negative reviews to construct dataset 2 for exploring tourists' demands. The statistical results of the ratio of positive and negative reviews in English reviews for each hotel in Dataset 2 are shown in **Figure 3**.

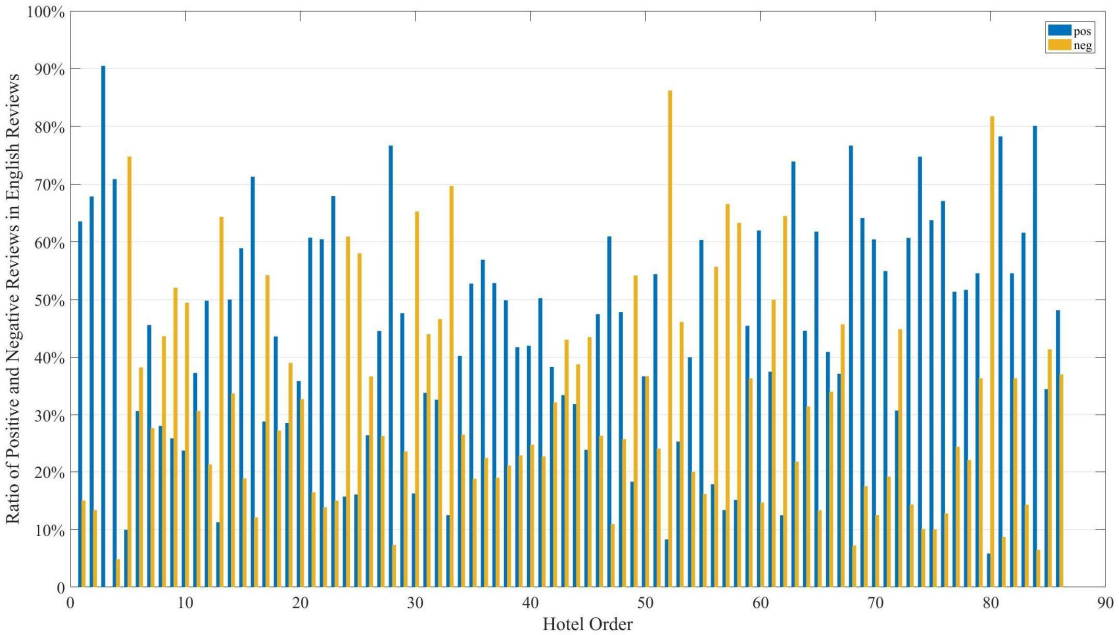

Figure 3. The ratio of positive and negative reviews in Dataset 2.

Figure 3 shows that the number of positive and negative reviews is roughly equal, indicating the rationality of the classification method regarding positive and negative reviews in this paper. In Dataset 2, the numbers of positive and negative reviews are 36,500 and 22,852, respectively.

2.3. Design of the study

English reviews (86 hotels) were selected for this paper. Based on all reviews, sentiment words in the field of tourist hotels were identified, and a sentiment dictionary was constructed. Additionally, based on positive and negative reviews, demand-oriented reviews were identified to explore tourist demands, providing references for the optimization of services in tourist hotels. On the one hand, this paper constructs a sentiment dictionary for the field of tourist hotels, providing references for research on text sentiment analysis in this domain. On the other hand, this paper designs effective methods to extract tourist demands from a large number of tourist reviews, thereby providing suggestions for service improvement in tourist hotels. The specific study design process and methodological rationale are as follows.

2.3.1. Identification of sentiment words and construction of the sentiment dictionary

Since sentiment words need to possess strong domain relevance [26], and currently, there are no sentiment dictionaries specific to the tourist hotel domain, this study constructs a sentiment dictionary based on textual data from the tourist hotel domain. This study identifies sentiment words based on PMI [29] and explores the boundaries of sentiment words based on IE [30], thereby recognizing sentiment words in the tourist hotel domain and constructing a sentiment dictionary to ensure that the segmentation results conform to this domain. Firstly, the principles of PMI and IE are introduced in (1) and (2), respectively, and then the sentiment word identification algorithm is described in (3).

(1) The method for discovering sentiment words based on PMI

The PMI measures the degree of mutual dependence between two words. In natural language processing, it is often used to identify new sentiment words [29]. This study calculates neighboring words' mutual information value to search for words with high mutual dependence. The probability of neighboring words with high mutual dependence forming a new word is greater. Finally, new words with high relevance to the research field are identified through manual analysis. Define PMI_{ij} as the PMI value between word i and its neighboring word j, with the calculation formula as follows.

$$PMI_{ij} = log\, P_{ij} / P_i \times P_j \quad j \in NW_i \tag{1}$$

where P_i and P_j represent the occurrence probabilities of words i and j in the text, P_{ij} represents the joint occurrence probability of words i and j in the text, and NW_irepresents the set of neighboring words of word i. Usually, when $PMI_{ij} > 0$, words i and j may form a new word; the larger the value of PMI_{ij}, the higher the likelihood of words i and j forming a new word. In practical usage, the neighboring words on the left and right are usually distinguished, and the pointwise mutual information values are calculated separately to determine the word formation direction

of new words. Define NW_i^L and NW_i^R as the sets of left and right neighboring words for word i respectively.

It is worth noting that when using PMI to identify sentiment words, it is necessary to impose restrictions on the minimum occurrence probability of words. If the occurrence probability of a word is too low, the word may be a "rare word," such as the personal name, place name, and specialized term, or a "misleading word," such as the punctuation marks, misspelled word, and uncommon spelling. Therefore, in practical usage, to improve the efficiency of sentiment word identification, restrictions should be placed on the occurrence probability of words, which is $P_i \geq p_0$, where p_0 is a constant parameter defined as "the minimum occurrence probability of words for sentiment word identification."

(2) The method for exploring the boundaries of sentiment words based on IE

Entropy originally refers to disorder. In the field of text sentiment analysis, word information entropy refers to the richness of the context in which a word is used, specifically the richness of the word's collocation with neighboring words [30]. The higher the degree of collocation with neighboring words, the greater the probability that a word is independent. In natural language processing, IE is often used to identify the boundaries of sentiment words. This method calculates the IE value of a word to determine whether the word is independent. Independent words form sentiment words and no longer form new words with neighboring words. The IE of word i is defined as E_i, with its calculation formula as follows.

$$E_i = -\sum_{j \in NM_i} (P_{ij}/P_i)\, log_2\, P_{ij}\, /P_i \tag{2}$$

The larger the value of E_i, the higher the probability that word i is an independent sentiment word. Let E^0 be the information entropy threshold for determining whether word i is an independent sentiment word. In practical usage, the neighboring words on the left and right are usually distinguished, and the left and right entropy are calculated separately. The left and right entropy of word i are defined as E_i^L and E_i^R, respectively, with their calculation formulas as follows.

$$E_i^L = -\sum_{j \in NM_i^L} (P_{ij}/P_i)\, log_2\, P_{ij}\, /P_i \tag{3}$$

$$E_i^R = -\sum_{j \in NM_i^R} (P_{ij}/P_i)\, log_2\, P_{ij}\, /P_i \tag{4}$$

It is worth noting that when determining the boundaries of new words, it is also necessary to impose restrictions on the minimum probability of word occurrence to exclude the influence of "rare words" and "misleading words." This is because, according to the description of "the minimum word occurrence probability for new word discovery" in (2) when the word occurrence probability is lower than p_0, it cannot form a new word, consistent with the concept of word independence. Therefore, the minimum probability threshold for determining independent words is also chosen as the threshold p_0.

(3) Identification of sentiment words and construction of sentiment dictionary based on PMI and IE

This study identifies sentiment words relevant to the research theme in the field

of tourist hotels and constructs a sentiment dictionary based on PMI and EI. The construction process iteratively follows these steps: segmenting the text into words and removing "stop words" which are meaningless words, such as punctuation marks, misspelled words, and uncommon spellings [38]; constructing the left and right neighboring word sets for each word; calculating the probability of each word appearing in the text and the joint occurrence probability of it with its neighboring words; computing the left and right information entropy for each word; determining whether a word is an independent word based on its left and right information entropy, and constructing a set of non-independent words; calculating the PMI value between each word in the non-independent word set and each neighboring word; sorting the PMI values of each word with its neighboring words from high to low, and determining the newly formed words based on the research theme; adding the newly formed words to the sentiment word set. The algorithm iterates until the set of non-independent words is empty, at which point the algorithm terminates. The pseudocode for this algorithm is as follows.

Algorithm 1. Identification of sentiment words

Input: Text set $\boldsymbol{T}$
Output: Sentiment word set $\boldsymbol{NW}$
Steps:

i. Preprocessing:

1) Load the stop word set $\boldsymbol{STW}$.
2) Load the minimum occurrence probability threshold p_0 and minimum EI threshold E^0 and for determining independent words.
3) Load the sentiment word set $\boldsymbol{NW}$ which is initially empty.

ii. Iterating:

1) Segment all texts in set $\boldsymbol{T}$, and check each segmentation to see if it is in set $\boldsymbol{STW}$. If it is, discard it; if not, keep it.
2) Calculate the occurrence probability of each segmentation of step ii.1) in set $\boldsymbol{T}$, forming vector $\boldsymbol{P}$.
3) Retrieve vector $\boldsymbol{P}$, filter out words with occurrence probabilities higher than threshold p_0, forming the candidate sentiment words set $\boldsymbol{WNW}$.
4) According to the segmentation results of step ii.1), identify the left and right neighboring words of each word in set $\boldsymbol{WNW}$, and construct the left neighboring word matrix $\boldsymbol{WNW\_L}$ and the right neighboring word matrix $\boldsymbol{WNW\_R}$.
5) Calculate the joint occurrence probabilities of each word in set $\boldsymbol{WNW}$ with its left and right neighboring words in set $\boldsymbol{T}$, forming matrices $\boldsymbol{P\_WL}$ and $\boldsymbol{P\_WR}$ respectively.
6) Calculate the left and right information entropy for each word in set $\boldsymbol{WNW}$.
7) Determine whether a word is an independent word based on its left and right information entropy. If both the left and right information entropies are not lower than threshold E^0, it is considered an independent word; otherwise, it is a non-independent word. Non-independent words form set $\boldsymbol{NIW}$.
8) Check whether set $\boldsymbol{NIW}$ is empty or not. If it is, terminate the iteration; if not, continue the iteration.
9) Perform the following steps for each word in set $\boldsymbol{NIW}$:
   A. Calculate the PMI value between the word and its neighboring words:
      a. If the left information entropy of the word is lower than threshold E^0 and the right information entropy is not lower than threshold E^0, then calculate the PMI value between the word and its left neighboring word.
      b. If the left information entropy of the word is not lower than threshold E^0 and the right information entropy is lower than threshold E^0, then calculate the PMI value between the word and its right neighboring word.

Algorithm 1. (*Continued*)

c. If both the left and right information entropies of the word are lower than threshold E^0, then calculate the PMI value between the word and its left and right neighboring words.

B. Sort the PMI values of each word in descending order and identify sentiment words based on the research theme.

C. Update set $\boldsymbol{NW}$: Add the identified sentiment words to set $\boldsymbol{NW}$.

10) Return to step ii.1).

iii. Return sentiment word set $\boldsymbol{NW}$.

2.3.2. Identification and classification of tourist demands

The ambiguity of texts makes it difficult to identify demands [27]. To address this issue, based on the characteristics of text data in the dataset, this study categorizes the types of reviews containing tourist demands and summarizes their characteristics, as shown in **Table 2**.

Table 2. Five types of reviews contain tourist demands and their characteristics.

Type	Characteristics
Positive reviews	Such reviews typically contain objects expressing positive sentiment and words expressing positive sentiment. The typical word combinations include noun + linking verb + adjective, adjective + noun, linking verb + adjective + noun.
Negative reviews	Such reviews typically contain objects expressing negative sentiment and words expressing negative sentiment. The typical word combinations include noun + linking verb/linking verb in negative form + adjective, adjective + noun, linking verb/linking verb in negative form + adjective + noun.
Suggestion reviews	Such reviews typically contain modal verbs such as "should," "could," "need," and "suggest," or words indicating suggestions such as "suggestion," "advice," "advise," and "propose." In this paper, these types of words are referred to as "suggestion words," forming the suggestion word list.
Demand reviews	Such reviews typically contain words expressing needs or demands, such as "want," "need," and "demand," or words indicating a lack of something needed, such as "lack." In this paper, these types of words are referred to as "demand words," forming the demand word list.
Comparison reviews	Such reviews typically contain comparative and superlative vocabulary or words used to introduce comparative sentence structures, such as "compared," "compare," and "rather than." In this paper, these types of words are referred to as "comparison words," forming the comparison word list.

According to the five types of reviews containing tourist demands and their characteristics, this study designs a method to identify and extract tourist demands for tourist hotels from reviews: (1) Constructing suggestion word list, demand word list, and comparison word list. (2) Segmenting reviews into single sentences, as most demands exist within a single sentence rather than spanning multiple sentences. (3) According to the summary in **Table 2** of the five types of reviews containing tourist demands and their characteristics, the form of each sentence is determined. If it matches one of the five forms, extract the demand accordingly. (4) Combining demands with the same meaning but different expressions and grouping similar demands to facilitate understanding. Additionally, to ensure the representativeness of demands, a minimum probability threshold for the occurrence of demands is imposed, assuming the minimum occurrence probability threshold for demand identification is p_1. The pseudocode for this algorithm is as follows.

Algorithm 2. Identification and classification of tourist demands

Input: Text set $\boldsymbol{T}$
Output: Demand set $\boldsymbol{D}$
Steps:
i. Preprocessing:
1) Load the stop word set $\boldsymbol{STW}$.
2) Load the suggestion word set $\boldsymbol{SW}$, demand word set $\boldsymbol{DW}$, and comparison word set $\boldsymbol{DW}$.
3) Load the minimum occurrence probability threshold p_1 for demand identification.
4) Load the demand set $\boldsymbol{D}$, which is initially empty.
ii. Segmenting all texts in set $\boldsymbol{T}$ into single sentences, all single sentences form set $\boldsymbol{SS}$.
iii. Check each sentence in set $\boldsymbol{SS}$ to identify its type according to **Table 2**, and identify demands from the single sentence based on its type and characteristics, put the demands into set $\boldsymbol{D}$.
iv. Calculate the occurrence probability of each demand in set $\boldsymbol{D}$ in set $\boldsymbol{T}$.
v. Check the occurrence probability of each demand in set $\boldsymbol{D}$ in collection $\boldsymbol{T}$: If it is lower than p_1, remove it from set $\boldsymbol{D}$; otherwise, keep it.
vi. Return the demand set $\boldsymbol{D}$.

3. Results and discussion

3.1. Identification of sentiment words and construction of sentiment dictionaries

Due to the different language conventions between review titles and review content, where review titles tend to be more concise and formal while review content is more specific and informal, this study applies the Identification Algorithm of Sentiment Words (Algorithm 1) separately to the data of review titles and contents to identify sentiment words and construct a sentiment dictionary. Section 3.1.1. details the parameter values of Algorithm 1 and the necessary preparations before its application, while Sections 3.1.2. and 3.1.3. present the results of sentiment word identification for review titles and contents.

3.1.1. Parameter assignment and preparation for Algorithm 1 application

The results of parameter assignment for Algorithm 1 are shown in **Table 3**.

Table 3. The results of parameter assignment for Algorithm 1.

Parameters	Meanings	Assigned values
p_0	the minimum occurrence probability threshold for determining independent words	0.003
E^0	the minimum EI threshold for determining independent words	1.5850

The tokenizer used in this paper is NLTK, which is usually used in English natural language processing [39]. Before applying Algorithm 1, it is necessary to construct a stop word list to remove "stop words" from the segmentation results, which are meaningless words such as punctuation marks, misspelled words, and uncommon spellings [38]. This study utilizes mainstream stop word lists, including CN stop words [38], HIT stop words [40], Baidu stop words [41], MIL-SCU stop words [42], and NLTK stop words [39]. Words of other languages and symbols are removed, duplicates are eliminated, and the lists are integrated. Additionally, through regular

expression matching, this study identifies segmented words containing sentiments and meaningless punctuation marks and replaces segmented words containing easily recognizable emoticons such as "$" (money), "❤" (love), "😵" (sad), etc., with their meanings; adds segmented words containing emoticons that are difficult to interpret, such as "😝," "😳," "🤗," etc., and meaningless punctuation marks to the stop word list. A total of 136 emoticons were identified and processed in this study, with 1,287 and 6,211 stop words identified for review titles and contents, respectively.

For the convenience of applying Algorithm 1, this study uniformly processes words into lowercase and transforms their forms according to the specific conversion methods shown in **Table 4**.

Table 4. Converted methods for words.

Original type	The type can be converted to	Examples
Words containing uppercase letters	Words entirely in lowercase letters	"Family" to "family"
Plural nouns	Singular formal	"families" to "family"
Third person singular of the verb	Verbs in infinitive form	"walks" to "walk"
Verbs in the present participle form	Verbs in infinitive form	"walking" to "walk"
Verbs in the past tense form	Verbs in infinitive form	"walked" to "walk"
Adjectives/Adverbs in comparative form	Adjectives/Adverbs in infinitive form	"friendlier" to "friendly"
Adjectives/Adverbs in superlative form	Adjectives/Adverbs in infinitive form	"best" to "good"

3.1.2. Identification of sentiment words for review titles

Applying Algorithm 1 to the review titles in Dataset 1 to identify sentiment words validates the algorithm's effectiveness in identifying sentiment words. The segmentation results before and after applying Algorithm 1 are compared. The top 60 segmented words by frequency are selected to generate a word cloud, visually presenting the segmentation results. In the word cloud, the font size represents the frequency of the words. The results are shown in **Figure 4**.

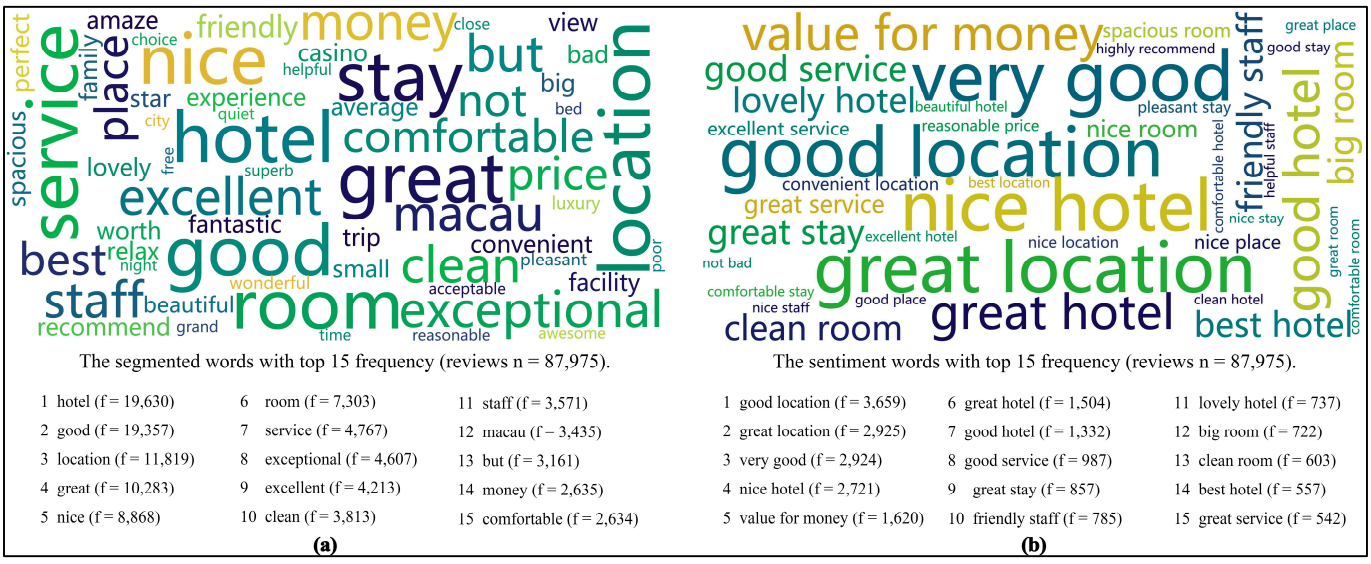

Figure 4. Comparison of segmentation results of review titles before and after identifying sentiment words using Algorithm 1 [Noted: **(a)** Before using Algorithm 1; **(b)** After using Algorithm 1].

The total number of sentiment words for review titles is 1,920. As shown in **Figure 4b**, the top 15 sentiment words extracted from the review titles are: "good location," "great location," "very good," "nice hotel," "value for money," "great hotel," "good hotel," "good service," "great stay," "friendly staff," "lovely hotel," "big room," "clean room," "best hotel," and "great service." From the results of sentiment word identification, it could be seen that tourists value the location, value for money, spaciousness and cleanliness of the room, attitude of staff, and service quality of the tourist hotel.

The comparison between **Figure 4a and 4b** reveals that after applying Algorithm 1, the segmentation results of review titles changed from single, ambiguous words to clear and meaningful word phrases. These contrast results demonstrate the effectiveness of Algorithm 1 in identifying sentiment words from review titles.

3.1.3. Identification of sentiment words for review contents

Applying Algorithm 1 to the review contents in Dataset 1, sentiment words are identified. The same validation and visualization methods as described in Section 3.1.2. are adopted, and the results are shown in **Figure 5**.

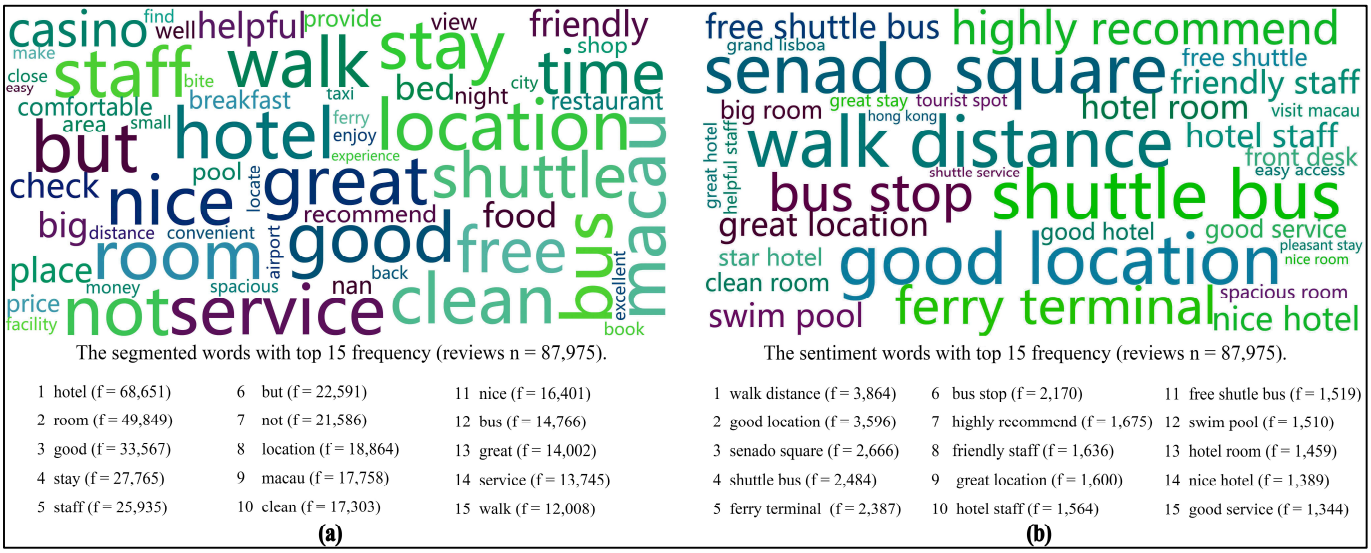

Figure 5. Comparison of segmentation results of review contents before and after identifying sentiment words using Algorithm 1 [Noted: **(a)** Before using Algorithm 1; **(b)** After using Algorithm 1].

The total number of sentiment words for review contents is 16,826. As shown in **Figure 5b**, the top 15 sentiment words extracted from the review contents are: "walk distance," "good location," "senado square" (a famous tourist attraction in Macau), "shuttle bus," "ferry terminal," "bus stop," "highly recommend," "friendly staff," "great location," "hotel staff," "free shuttle bus," "swim pool," "hotel room," "nice hotel," and "good service." From the results of sentiment word identification, it could be seen that tourists value the location, transportation convenience, distance between the hotel and tourist attractions, hotel facilities, room quality, attitude of staff, and service quality of the tourist hotel.

The comparison between **Figure 5a** and **5b** reveals that after applying Algorithm 1, the segmentation results of review contents changed from single, ambiguous words to clear and meaningful word phrases. These contrast results demonstrate the

effectiveness of Algorithm 1 in identifying sentiment words from review contents.

3.2. Identification and classification of tourist demands

The minimum occurrence probability threshold for determining independent words p_0 is set to 0.0002. Algorithm 2 is applied to Dataset 2, and similar demands are grouped. The results of demand identification and classification are shown in **Table 5**.

Table 5. Identification and classification of tourist demands.

Type	Demands contained (For Example)	Number of demands	Frequency of demands
Facility	"free wifi," "large bed," "hot water"	68	14,191
Face of the Room and Hotel	"clean room," "big room," "spacious and clean"	65	9,429
Transport	"10-minute walk," "shuttle bus provided, " "shuttle service"	64	25,386
Staff	"speak English," "nice and helpful," "hotel staff be friendly"	51	12,592
Food and Restaurant	"free breakfast," "free breakfast," "breakfast included"	29	5,622
Location	"city center," "convenient location," "good location"	25	9,744
Price	"worth to stay," "reasonable price," "worth money"	21	2,195
Nearby Tourist Attractions	"tourist destination," "tourist attraction," "Senado Square" (a well-known tourist attraction in Macao)	18	7,458
Room Type	"upgrade room," "family room," "twin room"	18	2,409
Smell and Smoking	"cigarette smell," "smoking smell," "bad smell"	12	2,338
View	"sea view," "city view," "nice view"	8	3,022
Save Time	"long queue," "wait for an hour," "fast check"	7	710
Easy to find	"easy to find," "easily accessible," "difficult to find"	5	210
Nearby Shops	"convenient to store," "shop center," "shop hotel"	5	770
All types	/	396	96,106

From **Table 5**, it could be seen that a total of 396 tourist demands were identified, with a total frequency of 96,106. These demands were ultimately divided into 14 categories, namely: "Facility," "Face of the Room and Hotel," "Transport," "Staff," "Food and Restaurant," "Location," "Price," "Nearby Tourist Attractions," "Room Type," "Smell and Smoking," "View," "Save Time," "Easy to find," "Nearby Shops." Among them, "Facility," "Face of the Room and Hotel," "Transport," "Staff," and "Food and Restaurant" are the top five demands most valued by tourists. Based on the analysis and existing research results, the impact of these five types of demands on tourist satisfaction could be analyzed, and corresponding recommendations could be proposed for tourist hotels.

"Facility": High-frequency demands in this category include "hot water," "free Wi-Fi," and "large and comfortable bed." After a day of sightseeing, tourists are often tired and have a high demand for facilities that help them relax. If tourist hotels could provide facilities that correspond to the needs of tourists, it could help them relax and relieve fatigue, thus allowing them to enjoy their trip better [43].

"Face of the Room and Hotel": The high-frequency demands in this category include "spacious and clean," "clean and tidy," and "comfortable and clean." With

excitement and happiness, tourists head to an unfamiliar tourist destination. If the tourist hotel could provide a clean and tidy room, it could make the tourists happier, thus enhancing their travel experience [44].

"Transport": The high-frequency demands in this category include "provide the shuttle bus," "bus service," and "ferry terminal hotel." Convenient and inexpensive transportation saves customers' travel time, effectively reduces customer anxiety, and increases travel flexibility. If the tourist hotel could provide convenient transportation services for customers, it could greatly enhance the travel experience of tourists [45].

"Staff": The high-frequency demands in this category include "helpful and polite," "helpful and friendly," and "can speak English." These demands mainly focus on the attitude and skills of the staff. If the staff of the tourist hotel could provide professional and friendly service, it could greatly enhance the tourists' positive perception of the hotel and improve their travel experience [46].

"Food and Restaurant": The high-frequency demands in this category include "provide free breakfast," "nearby Chinese restaurant," and "nice food." Food is also an important part of travel, as many tourists visit a new place for its cuisine [47]; additionally, free breakfast could save tourists a lot of time during their travels. If the hotel could provide breakfast and delicious food for tourists as much as possible, it could greatly enhance their travel experience [47].

4. Conclusion

In recent years, with the gradual improvement of people's living standards around the globe, tourists' demands and standards for tourist hotel construction have been rising accordingly. Furthermore, understanding tourists' demand reviews and extracting core viewpoints from consumer reviews are crucial for optimizing hotel services and enhancing tourist satisfaction. For the service quality improvement of tourist hotels, this paper takes Macao, a famous tourist destination in China, as an example, based on tourist reviews of tourist hotels on the Agoda platform, to explore the demands of tourists in tourist hotels from the perspective of text sentiment analysis.

The present study designs Algorithm 1 based on PMI and IE to identify sentiment words in the field of tourist hotels and constructs a sentiment dictionary. The comparison (**Figures 4** and **5**) of segmentation results of reviews before and after identifying sentiment words using Algorithm 1 demonstrates the effectiveness of Algorithm 1 in identifying sentiment words from review contents. This aims to address the problem of poor relevance between word segmentation results and the field of tourist hotels due to the lack of a related sentiment dictionary. Afterward, this study designs Algorithm 2 based on the summarization of the five types of reviews containing tourist demands and their characteristics to solve the problem of ambiguity of texts and accurately identify tourist demands and group similar tourist demands into the same categories to address the problem of multiple expressions for the same demand. Finally, demands are categorized into 14 categories. This study selected the five demand types, "Facility," "Face of the Room and Hotel," "Transport," "Staff," and "Food and Restaurant," with the highest frequencies to analyze their importance for tourist satisfaction and according to the demands to provide suggestions for tourist hotels to improve their service.

In summary, this paper identifies sentiment words in the field of tourist hotels. It constructs a sentiment dictionary, providing a reference for research on sentiment analysis of text in the field of tourist hotels. Besides, it accurately identifies and groups tourist demands to provide directions for the sustainable optimization improvement of tourist services. For future research by other scholars in this field, this study suggests that more related platforms' text data could be collected to build a more systematic and comprehensive sentiment dictionary. More expressions of demand could be explored from consumers' comments to satisfy the service providers, improve their service level, and build a sustainable business service environment together.

Author contributions: Conceptualization, LW, XZ and JX; methodology, LW, XZ and JX; software, LW, HZ and ZZ; validation, LW and HZ; formal analysis, LW and XZ; investigation, LW, XZ and CZ; resources, LW and XZ; data curation, LW and XZ; writing—original draft preparation, LW and XZ; writing—review and editing, LW, HZ and JX; visualization, LW, HZ, JX and CZ; supervision, HZ and JX; project administration, JX; funding acquisition, JX. All authors have read and agreed to the published version of the manuscript.

Data availability statement: All the data for this study is available upon request to the author.

Conflict of interest: The authors declare no conflict of interest.

References

1. Pham Minh Q, Ngoc Mai N. Perceived risk and booking intention in the crisis of COVID-19: comparison of tourist hotels and love hotels. Tourism Recreation Research. 2021; 48(1): 128-140. doi: 10.1080/02508281.2021.1885798
2. Gössling S, Lund-Durlacher D. Tourist accommodation, climate change and mitigation: An assessment for Austria. Journal of Outdoor Recreation and Tourism. 2021; 34: 100367. doi: 10.1016/j.jort.2021.100367
3. Wang M, Liu J, Zhang S, et al. Spatial pattern and micro-location rules of tourism businesses in historic towns: A case study of Pingyao, China. Journal of Destination Marketing & Management. 2022; 25: 100721. doi: 10.1016/j.jdmm.2022.100721
4. Chiawo D, Haggai C, Muniu V, et al. Tourism recovery and sustainability post pandemic: An integrated approach for Kenya's tourism hotspots. Sustainability. 2023; 15(9): 7291. doi: 10.3390/su15097291
5. Liu X, Abhari K, Wang W. Resurgence in paradise: decoding the patterns of arrivals with different trip purposes in Hawaii's post-pandemic tourism recovery. Current Issues in Tourism. 2023; 1-7. doi: 10.1080/13683500.2023.2277903
6. Market Research Excellence Center. Hotel and other travel accommodation market size, shaping future trends and growth from 2023-2030. Available online: https://www.linkedin.com/pulse/hotel-other-travel-accommodation-market-size-share-iwbxf (accessed on 15 April 2024).
7. Liao W, Wang H, Xu J. The spatial structure characteristic and road traffic accessibility evaluation of A-level tourist attractions within Wuhan Urban Agglomeration in China. 3C Tecnología. 2023; 12(2): 388-409. doi: 10.17993/3ctecno.2023.v12n3e45.388-409
8. Statista. Hotels-Worldwide. Available online: https://www.statista.com/outlook/mmo/travel-tourism/hotels/worldwide (accessed on 16 April 2024).
9. Srivastava P, Mishra N, Singh N, et al. Beyond carbon footprints: the 'Greta Thunberg Effect' and tourist hotel preferences. Journal of Travel & Tourism Marketing. 2024; 41(4): 578-595. doi: 10.1080/10548408.2023.2293017
10. Chan ESW, Wong SCK. Hotel selection: When price is not the issue. Journal of Vacation Marketing. 2006; 12(2): 142-159. doi: 10.1177/1356766706062154

11. Iversen EK, Holmen RB. The tourism industry in Vestland during the green transition: Stakeholder perspectives on challenges and opportunities. Available online: https://vista-analyse.no/site/assets/files/8133/snf_07_23.pdf (accessed on 15 April 2024).
12. Vives A, Jacob M. Sources of price elasticity of demand variability among Spanish resort hotels: a managerial insight. Journal of Hospitality and Tourism Technology. 2023; 14(2): 137-153. doi: 10.1108/jhtt-11-2020-0298
13. Wang C, Hao Y. Empirical analysis of tourist satisfaction of leisure farms: evidence from Qing Jing Farms, Taiwan. Humanities and Social Sciences Communications. 2023, 10: 384. doi: 10.1057/s41599-023-01901-w
14. Wu C, Cui J, Xu X, et al. The influence of virtual environment on thermal perception: physical reaction and subjective thermal perception on outdoor scenarios in virtual reality. International Journal of Biometeorology. 2023; 67(8): 1291-1301. doi: 10.1007/s00484-023-02495-3
15. Wu C. The impact of public green space views on indoor thermal perception and environment control behavior of residents - A survey study in Shanghai. European Journal of Sustainable Development. 2023; 12(3): 131. doi: 10.14207/ejsd.2023.v12n3p131
16. Park JH, Kim NJ. Influence of hotel guests' perception of ESG management importance on their willingness to accept losses through perceived value, trust, emotional well-being, and brand preference. The Tourism Sciences Society of Korea. 2023; 47(7): 69-90. doi: 10.17086/jts.2023.47.7.69.90
17. Patwary AK, Aziz RC, Hashim NAAN. Investigating tourists' intention toward green hotels in Malaysia: a direction on tourist sustainable consumption. Environmental Science and Pollution Research. 2022; 30(13): 38500-38511. doi: 10.1007/s11356-022-24946-x
18. Chen G, Liu M, Zhang Y, et al. Using images to detect, plan, analyze, and coordinate a smart contract in construction. Journal of Management in Engineering. 2023; 39(2). doi: 10.1061/jmenea.meeng-5121
19. Liu S, Li X, He C. Study on dynamic influence of passenger flow on intelligent bus travel service model. Transport. 2021; 36(1): 25-37. doi: 10.3846/transport.2021.14343
20. He C, Liu M, Hsiang, SM, et al. Synthesizing ontology and graph neural network to unveil the implicit rules for US bridge preservation decisions. Journal of Management in Engineering. 2024; 40(3). doi: 10.1061/JMENEA.MEENG-5803
21. Swasto DF, Rahmi DH, Rahmawati Y, et al. Proceedings of the 6th International Conference on Indonesian Architecture and Planning (ICIAP 2022). Springer Nature Singapore; 2023. doi: 10.1007/978-981-99-1403-6
22. Matrutty JP, Adrian AM, Angdresey A. Sentiment analysis of visitor reviews on star hotels in Manado City. Journal of Information Technology and Computer Science. 2023; 8(1): 21-32. doi: 10.25126/jitecs.202381403
23. Yu W, Cui F, Hou Z. The evolution of consumers' demand for hotels under the public health crisis: opinion mining from online reviews. Current Issues in Tourism. 2022; 26(12): 1974-1990. doi: 10.1080/13683500.2022.2073204
24. Çevrimkaya M, Çavus Ş, Şengel Ü. Assessment of hotels' online complaints in domestic tourism: mixed analysis approach. International Journal of Tourism Cities. 2024. doi: 10.1108/ijtc-01-2023-0007
25. Çelik MN, Çevirgen A. The role of accommodation enterprises in the development of sustainable tourism. Journal of Tourism and Services. 2021; 12(23): 181-198. doi: 10.29036/jots.v12i23.264
26. Ahmed M, Chen Q, Li Z. Constructing domain-dependent sentiment dictionary for sentiment analysis. Neural Computing and Applications. 2020; 32(18): 14719-14732. doi: 10.1007/s00521-020-04824-8
27. Yin F, Wang Y, Liu J, et al. The construction of sentiment lexicon based on context-dependent part-of-speech chunks for semantic disambiguation. IEEE Access. 2020; 8: 63359-63367. doi: 10.1109/access.2020.2984284
28. Alrasheed H. Word synonym relationships for text analysis: A graph-based approach. PLOS ONE. 2021; 16(7): e0255127. doi: 10.1371/journal.pone.0255127
29. Ahanin Z, Ismail MA. A multi-label emoji classification method using balanced pointwise mutual information-based feature selection. Computer Speech & Language. 2022; 73: 101330. doi: 10.1016/j.csl.2021.101330
30. Wang Z, Wang L, Ji Y, et al. A novel data-driven weighted sentiment analysis based on information entropy for perceived satisfaction. Journal of Retailing and Consumer Services. 2022; 68: 103038. doi: 10.1016/j.jretconser.2022.103038
31. Du Z, Huang AG, Wermers R, et al. Language and domain specificity: A Chinese financial sentiment dictionary. Review of Finance. 2021; 26(3): 673-719. doi: 10.1093/rof/rfab036
32. Huang CH, Tsaur JR, Yang CH. Does world heritage list really induce more tourists? Evidence from Macau. Tourism Management. 2012; 33(6): 1450-1457. doi: 10.1016/j.tourman.2012.01.014

33. Chu CL. Spectacular Macau: Visioning futures for a world heritage city. Geoforum. 2015; 65: 440-450. doi: 10.1016/j.geoforum.2015.06.009
34. Ung A, Vong TN. Tourist experience of heritage tourism in Macau SAR, China. Journal of Heritage Tourism. 2010; 5(2): 157-168. doi: 10.1080/17438731003668502
35. Araújo M, Pereira A, Benevenuto F. A comparative study of machine translation for multilingual sentence-level sentiment analysis. Information Sciences. 2020; 512: 1078-1102. doi: 10.1016/j.ins.2019.10.031
36. Jain DK, Boyapati P, Venkatesh J, et al. An intelligent cognitive-inspired computing with big data analytics framework for sentiment analysis and classification. Information Processing & Management. 2022; 59(1): 102758. doi: 10.1016/j.ipm.2021.102758
37. Yadollahi A, Shahraki AG, Zaiane OR. Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys. 2017; 50(2): 1-33. doi: 10.1145/3057270
38. Wu M, Jiang T, Bu C, et al. Coarse-to-fine entity alignment for Chinese heterogeneous encyclopedia knowledge base. Future Internet. 2022; 14(2): 39. doi: 10.3390/fi14020039
39. Sarica S, Luo J. Stopwords in technical language processing. PLOS ONE. 2021; 16(8): e0254937. doi: 10.1371/journal.pone.0254937
40. Xin Y, Tan X, Ren X. Will the relaxation of COVID-19 control measures have an impact on the Chinese internet-using public? Social media-based topic and sentiment analysis. International Journal of Public Health. 2023; 68. doi: 10.3389/ijph.2023.1606074
41. Xu A, Qi T, Dong X. Analysis of the Douban online review of the MCU: based on LDA topic model. Journal of Physics: Conference Series. 2020; 1437(1): 012102. doi: 10.1088/1742-6596/1437/1/012102
42. Zhong B, He W, Huang Z, et al. A building regulation question answering system: A deep learning methodology. Advanced Engineering Informatics. 2020; 46: 101195. doi: 10.1016/j.aei.2020.101195
43. Bahar AM, Maizaldi M, Putera N, et al. The effect of tourism facilities, service quality and promotion of tourist satisfaction in South Pesisir District. J-MAS (Jurnal Manajemen dan Sains). 2020; 5(1): 5. doi: 10.33087/jmas.v5i1.141
44. Wirakusuma RM, Samudra AAA, Sumartini NK. Investigating the impact of sensory experience on budget hotel rooms to maximize guests satisfaction. Journal of Business on Hospitality and Tourism. 2021; 7(1): 171. doi: 10.22334/jbhost.v7i1.284
45. Virkar AR, Mallya PD. A review of dimensions of tourism transport affecting tourist satisfaction. Indian Journal of Commerce & Management Studies. 2018; 9(1): 72-80. doi: 10.18843/ijcms/v9i1/10
46. Han J, Zuo Y, Law R, et al. Service quality in tourism public health: Trust, satisfaction, and loyalty. Frontiers in Psychology. 2021; 12. doi: 10.3389/fpsyg.2021.731279
47. Kim S, Choe JY, Kim PB. Effects of local food attributes on tourist dining satisfaction and future: The moderating role of food culture difference. Journal of China Tourism Research. 2020; 18(1): 121-143. doi: 10.1080/19388160.2020.1805667