Analysis of the Influence of Text Features on the Usefulness of Information: A Case of Tourism Text

Wen-xiao Jiang¹ and Ruo-yu Song²

¹,²School of Economics and Management
Chongqing University of Posts and Telecommunications
2, Chongwen Road, Nan’an District, Chongqing-400065, China.
²Email: 451417524@qq.com
*Corresponding author. *Email: 642021025@qq.com
Received 1 December 2022; Accepted 24 December 2022

Abstract. This paper takes 9260 domestic free tourism strategy data collected from the Mafengwo website as samples, quantifies the travel guide texts by using data mining, and performs word frequency statistics and keyword extraction on the data through Python code and NLPIR platform, then divides the high-weight words into three categories that affect the usefulness of information through hierarchical clustering. Ten hypotheses are proposed in the paper based on previous research. And a negative binomial regression model is built to conduct analysis. The results show that when the number of reads is regarded as a control variable, all ten text features, such as the rate of containing pictures, have a significant positive relationship on information usefulness. Therefore, suggestions are provided to develop influential and high-quality tourism strategies in terms of text features.

Keywords: text features, information usefulness, hierarchical clustering, negative binomial regression

1. Introduction

In recent years, the rapid development of Internet technology has injected new vitality into traditional industries, among which, the integration of tourism and the Internet has led to the rise of online travel platforms with the Internet as the carrier, through which people can make travel reservations and view tourism strategy. According to the 49th China Internet Development Statistics Report released by CNNIC in Beijing, as of December 2021, the size of China's Internet users reached 1.032 billion, and the Internet penetration rate reached 73.0%; among them, the size of travel texts users reached 397 million, accounting for 38.5% of the overall Internet users. Tourism strategy as a kind of information product based on the sharing of massive users' journey experiences and travel information are favored by more and more travelers, and more and more people are willing to share their tourism strategy with others when they go out to travel, thus generating a large number of related texts. tourism strategy are UGC(User Generated Content) products for the travel market, usually written by core users or officials, with the main purpose of recording users' information about various aspects of travel. It is a comprehensive evaluation of the whole
The information conveyed by the tourism strategy on the platform will influence users' judgment and decision of journey. The excellent quality of tourism strategy will attract more users to join the tips community on the platform, which will drive earnings growth in travel-related products. Secondly, it provides suggestions for UGC users to write tourism strategy, so as to better record and share their travel experiences, and provide suggestions and references for other users. In addition to tourists, some tips are also written by merchants, who provide users with tourism strategy, tweets and other information to reduce their perceived risk of consumption and thus influence their consumption decisions. Moreover, because the accuracy and usefulness of platform tourism strategy are affected by the subjective factors of their authors, there are often situations that do not reflect the real viewing experience, and the accumulation of tourism strategy on the platform is large, which makes it difficult for users to choose. Therefore, the textual features of tourism strategy and how such features affect the usefulness of information is a topic worth studying.

2. Literature review
Throughout the previous studies, Chatterjee first introduced the concept of online review usefulness, i.e., the degree of impact of the use of review information [1]. Online reviews are a type of online text. Mudambi et al. were the first to define online review usefulness in terms of perceived value, which is a measure of the extent to which consumers adopt online review information in their purchase decisions and the perceived value of it [2]. Since this perception-based value judgment is subjective in nature, it cannot accurately measure the usefulness of textual information. Internet Word of Mouth (IWOM), which mostly originates from the real experiences of travelers, provides unbiased, highly reliable, non-commercial information [3]. And tips are a type of word-of-mouth. Kuo found that user word-of-mouth can influence travel purchase intentions [4]. Litvin found that word-of-mouth can influence the hotel and tourism industry to some extent, and online platforms have become a key channel for travelers and word-of-mouth communicators to analyze and understand information about tourism products, which is similar to a short sharing of travel experiences [5].

Liang found that features such as the number of pictures of qunar.com hotel reviews showed a significant positive effect on usefulness, while review affective tendency and outward centrality had a significant negative effect, and review feature factors had a greater overall effect on review usefulness than reviewer feature factors [6]. Yang Xin Using multiple regression analysis, we obtained that review validity, review rate with graph, merchant response rate and merchant response length significantly and positively affect the sales of tourism products, and review variance significantly and negatively affects the sales of tourism products [7]. Wang constructed a framework model of factors influencing the usefulness of ELM online reviews, and the results showed that review information characteristics (central path) in review length and review volume positively influenced online review usefulness, extreme reviews (positive or negative reviews) were more useful than neutral reviews, and the stronger the reviewer's professionalism in reviewer characteristics (marginal path), the higher the review usefulness [8].

Based on the study of text features, some scholars have gone further to analyze online tourism texts by building topic models and establishing semantic network analysis maps.
through machine learning techniques. They focus their research on analyzing the subjective feelings and preferences of tourists reflected in the textual information as well as providing suggestions for case tourism places.

Wu A detailed analysis of UGC travel guide texts was conducted through a case study method using the Mafengwo website as an object to explore the text generation model [9]. Maria explored tourists’ experiences related to cultural attractions, similarities and differences among cultural attractions, and tourists’ preferences by conducting content analysis and principal component analysis on Italian online review texts posted on TripAdvisor by visitors to 58 cultural attractions in Naples (Italy) [10]. Taecharungroj used machine learning techniques to examine the online review texts of various tourist attractions in Phuket, Thailand, to suggest various aspects for the Tourism Authority of Thailand (TAT) to improve the attractions [11].

3. Text Analysis

3.1. Text selection

In terms of text selection, this paper selects the text of Mafengwo website as the research object. Mafengwo website (https://www.mafengwo.cn/) is a leading travel platform with a huge amount of tourism strategy texts in China. Users can express their opinions and favorites on the guide content by commenting, collecting and forwarding. In addition, Mafengwo’s strict requirements on the output make its strategy texts of high quality. Texts on this platform are relatively high quality travel texts.

This paper uses the crawler written by Python to crawl the page data of 34 provincial administrative regions of China's free tourism strategy on Mafengwo website, and get 9739 tips data, each of which can get its title, release time, number of experiences, number of views, number of reads, links, number of comments, number of favorites, number of retweets, number of recommendations, author's name, whether it is To ensure the integrity and reliability of the data as well as the results, data with text length less than 200 and collection number less than 5 were excluded from this study, leaving a total of 9260 valid text data.

3.2. Word segmentation processing

This paper analyzes the usefulness of text features for text information by quantifying the collected text for data mining and extracting the feature words that contain effective information as the database. Because a large number of ideographic numbers in the crawled text will be redundant and form "noise" after the word separation operation, the numbers in the text are first filtered out and deleted.

Unlike English text which can be segmented by words, there are no obvious separators in Chinese statements to segment words. After word segmentation computer can identify the meaning of the sentence more effectively, so the Chinese text data must be segmented. Jieba library can perform word segmentation, word annotation, keyword extraction and other functions on Chinese text, and support custom dictionaries. After using Jieba for word segmentation, there will be a large number of deactivation words, i.e. words that appear very frequently in the text retrieval process but have no relevant content, which need to be removed. In this paper, after updating and expanding the table of commonly used Chinese deactivated words, the deactivated words are removed according to the table and the result
3.3. Statistical analysis of word frequency

Word frequency is an indicator of the importance of a single word. In this paper, we use Python code to merge all documents and perform word frequency statistics, both to count the number of occurrences of a given word in a unit document and to filter the top 100 high-frequency words into the document.

Referring to the categories in the conceptual model of factors influencing tourist satisfaction proposed by Song Mingzhen and its research method about word frequency [12]. In order to analyze the degree of influence of different factors, the top 100 high-frequency words in the guide were matched with the categories related to tourism respectively. The analysis of the statistical results shows that the high-frequency words in the text mainly focus on the categories of scenic spot introduction, tour program, scenic spot characteristics, itinerary, emotional evaluation, time schedule, author’s suggestion, food, transportation, price discount, accommodation, supporting services, pictures, etc.

3.4. Cluster analysis

Clustering refers to the process of dividing a given set of objects into different subsets, with the goal of making the elements within each subset as similar as possible and the elements between different subsets as dissimilar as possible. The initial partitioning of raw data by clustering can improve the macroscopic understanding of the data. From a statistical point of view, cluster analysis is a way to simplify data through data modeling. In this paper, we adopt the structure of hierarchical clustering, with a top-down coalescent method for clustering.

In addition, the TextRank algorithm was used to extract the keywords of each text, and then the top 70 words ranked by the number of keywords were selected for hierarchical class clustering, and the results of the hierarchical clustering tree diagram were generalized, and the clustering results were summarized into three aspects: scenic spot introduction, itinerary arrangement and transportation, and diet (as shown in Figure 1), which is roughly consistent with the previous word frequency statistics and keyword extraction results, and can more accurately reflect the characteristics of tourism texts, indicating that the introduction of scenic spots, itinerary arrangement and transportation and diet in tourism texts can have a certain influence on the usefulness of information.

![Figure 1: Simplified diagram of hierarchical clustering](image)

By looking in the dictionary, it was found that these seventy words contain dictionary
words for money, space, movement, time, degree adverbs, perceptual experience and emotional experience, etc. Also, because the pictures inserted in the travel strategy will indicate the source, the above-mentioned types of words and picture factors can be used as text feature factors that affect the usefulness of information in conducting empirical studies.

4. Research model
4.1. Research hypothesis
Based on the results obtained from the data analysis above, the textual features of the tourism texts were summarized into three main dimensions, namely, introduction to scenic spots, itinerary and transportation, and food and drink. The relevant descriptions commonly used in real texts were then selected as specific measurement variables for the above three dimensions to meet the needs of combining realized texts. The final selection of figure-containing rate, spatial word rate, perceptual experience word rate and emotional experience word rate can reflect the introduction of scenic spots; the number rate, movement word rate and time word rate can reflect the trip arrangement and transportation; the figure-containing rate, money word rate and degree adverb rate can reflect the food and drink situation.

In this paper, we focus on the relationship between the influence of text features on the usefulness of information in 10 aspects, such as the rate of image content.

(1) Picture rate refers to the ratio of the number of pictures to the length of the text. Generally speaking, the strategy containing pictures is richer and more realistic than pure text, and the information conveyed can effectively enhance users' desire to collect. Zhang et al. showed that uploading images had a positive effect on the usefulness of reviews [14]. Wang and Chen proposed that online reviews with a combination of images and text would increase consumers' perceived usefulness compared to online reviews with only images and only text [15].

\[ H_1: \text{Picture rate has a positive effect on the usefulness of information.} \]

(2) Text length refers the number of characters in the content of the travel guide. Text length is closely related to the amount of information contained in the text, usually the longer the text length, the more valuable information contained. Many studies have shown that the length of the review has a positive effect on the usefulness of the review. Therefore, the hypothesis is proposed.

\[ H_2: \text{Text length has a positive effect on information usefulness.} \]

(3) Number rate is the ratio of the total number of Arabic and Chinese numbers in the guide to the number of words in the article. The numerical rate can visually reflect the proportion of detailed data about the tour expenses and travel distance in the guide. Generally speaking, the text with high numerical rate looks more intuitive, professional and reliable.

\[ H_3: \text{Number rate has a positive effect on the usefulness of information.} \]

(4) Money word rate is defined as the ratio of the number of money words in the guide to the number of words in the article. The price, discount and transportation cost in the guide are important information for the viewers. The money word rate can reflect the proportion of the cost of the tour introduced in the guide.

\[ H_4: \text{Money word rate has a positive effect on information usefulness.} \]

(5) Spatial word rate is the ratio of the number of spatial words in the strategy to the
number of words in the article. The text content contains a large amount of location-related information such as location of attractions, location nouns, geographic relationship nouns and location labels, etc [16]. The spatial word rate reflects the number of words in the guide. The rate of spatial words can reflect the proportion of scenic spots and places to play in the guide.

H5: Spatial word rate has a positive effect on information usefulness.

6) Mobile word rate is defined as the ratio of the number of mobile words in the guide to the number of words in the article. The meaning of move is to change the original position, the text content of the guide contains the traffic of tourism, travel mode and other mobile information [16]. The rate of mobile words reflects the number of words in the guide. The rate of mobile words can reflect the proportion of the introduction of transportation in the guide.

H6: Mobile word rate has a positive effect on information usefulness.

7) Time word rate is the ratio of the number of time words in the guide to the number of words in the article. The content of the text contains a lot of time-related information such as departure time, visiting time, scenic opening time and number of days of travel, etc [16]. The time word rate reflects the number of words in the guide. The time word rate can reflect the proportion of the introduction of scenic spots such as opening hours and tour schedule in the guide.

H7: Time word rate has a positive effect on information usefulness.

(8) Degree adverb rate is the ratio of the number of degree adverbs in the strategy to the number of words in the text. Generally speaking, travel texts frequently use degree adverbs to emphasize the degree of superiority of the described object in a certain aspect and their own feelings, such as "slightly", "very", "especially", etc [17]. The rate of degree adverbs can be somewhat higher. The rate of degree adverbs can reflect to some extent the intensity of the author's emotions and the level of detail in the description. Many scholars also consider the impact of degree adverbs when studying the usefulness of information.

H8: Degree adverbial rate has a positive effect on information usefulness.

(9) Perceived experience word rate is the ratio of the number of perceived experience words in the guide to the number of words in the article. Travel guides provide objective information while using a lot of ink for subjective descriptions. The perceptual experience lexicon contains a large number of intuitive sensations of what one sees and hears during the tour such as scenery, weather, temperature, food taste, etc. It is mainly used to analyze the perception of the tourist's cognitive image of the tourist destination. The perceptual dimension is the main way for scholars to measure and evaluate the image of the destination. The perceptual experience word rate can reflect the percentage of travel experiences and feelings of the authors in the guide.

H9: Perceived experience word rate has a positive effect on information usefulness.

(10) Emotional experience word rate is the ratio of the number of emotional experience words in the strategy to the number of words in the article. Travelers in the process of travel will be constantly stimulated by the new things from the outside world, the tourist's psychology will also change and produce positive or negative emotions such as pleasure, satisfaction, loss, pity, etc., and various emotions constitute the complete journey of the tourist [18]. The various emotions constitute the tourist's complete travel experience. Emotional experience word rate can reflect the proportion of the author's emotional evaluation and tour experience in the guide. Text sentiment analysis based on
sentiment dictionaries is a more widely used sentiment analysis method, and many scholars also use sentiment dictionaries for sentiment analysis of various texts to explore user satisfaction.

H10: Emotional experience word rate has a positive effect on information usefulness.

4.2. Variable indicators
Sussman et al. proposed the Information Adoption Model (IAM), which states that information quality and source reliability affect the usefulness of information, and thus the adoption of information [19]. This suggests that users judge the value of information before making information adoption. Information quality reflects the influence of textual characteristics, while source reliability reflects the influence of the source characteristics of the author of the strategy. The credibility of the source reflects the influence of the author's characteristics. Most of the strategy texts used in this paper is certified merchants and authors of Mafengwo website, and the information sources are reliable, so this paper focuses on text features in the information quality dimension. The number of favorites is used as the dependent variable to measure the usefulness of the information. The number of favorites, spatial word rate, perceptual experience word rate, emotional experience word rate, numerical rate, mobile word rate, time word rate, figure rate, money word rate and degree adverb rate are used as independent variables for verification. In addition to the factors considered above, other important factors also have an impact on the number of collections of strategy, such as the number of reads. Therefore, in this paper, the number of reads is used as a control variable to control its effect on the number of collections.

4.3. Statistical descriptive analysis
Among the variables needed in the hypothesis of factors influencing the usefulness of text features on information are: number of collections, picture rate, text length, numerical rate, money word rate, spatial word rate, mobile word rate, time word rate, degree adverb rate, perceptual experience word rate, emotional experience word rate. The above variables were obtained by Python code analysis to calculate the text data. Descriptive statistical analysis of the above variables was performed using SPSS 26 software, and the results are shown in Table 1.

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Minimum value</th>
<th>Maximum value</th>
<th>Average value</th>
<th>Standard deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture Rate</td>
<td>9260</td>
<td>0</td>
<td>0.01515</td>
<td>0.00078</td>
<td>0.001009</td>
<td>0</td>
</tr>
<tr>
<td>Text Length</td>
<td>9260</td>
<td>202</td>
<td>71428</td>
<td>3942.58</td>
<td>2866.199</td>
<td>8215099.5</td>
</tr>
<tr>
<td>Number rate</td>
<td>9260</td>
<td>0.011</td>
<td>0.336</td>
<td>0.07972</td>
<td>0.025051</td>
<td>0.001</td>
</tr>
<tr>
<td>Degree adverbs Rate</td>
<td>9260</td>
<td>0.005</td>
<td>0.290</td>
<td>0.07592</td>
<td>0.020808</td>
<td>0</td>
</tr>
<tr>
<td>Money word rate</td>
<td>9260</td>
<td>0.000</td>
<td>0.265</td>
<td>0.02457</td>
<td>0.018965</td>
<td>8.34E+10</td>
</tr>
<tr>
<td>Spatial word rate</td>
<td>9260</td>
<td>0.094</td>
<td>0.590</td>
<td>0.28080</td>
<td>0.053945</td>
<td>0.003</td>
</tr>
<tr>
<td>Mobile word rate</td>
<td>9260</td>
<td>0.010</td>
<td>0.348</td>
<td>0.09386</td>
<td>0.028499</td>
<td>0.001</td>
</tr>
<tr>
<td>Time word rate</td>
<td>9260</td>
<td>0.009</td>
<td>0.293</td>
<td>0.09388</td>
<td>0.025709</td>
<td>0.001</td>
</tr>
<tr>
<td>Perceptual experience</td>
<td>9260</td>
<td>0.016</td>
<td>0.304</td>
<td>0.09919</td>
<td>0.029473</td>
<td>0.001</td>
</tr>
</tbody>
</table>
4.4. Regression model and results

According to the results of the statistical descriptive analysis, we can see that the standard deviation of the dependent variable collection number is 4325.147 and the variance is 18706899.9, which is much larger than its mean value of 1891.49. The sample data are excessively discrete, and combined with the fact that the dependent variable is a count variable, the dependent variable is therefore suitable for analysis using a negative binomial regression model. The negative binomial regression model constructed in this paper is as follows.

\[ \text{Number of collections} = \beta_1 \times \text{Picture rate} + \beta_2 \times \text{Text Length} + \beta_3 \times \text{Number Rate} + \beta_4 \times \text{Money words rate} + \beta_5 \times \text{Spatial word rate} + \beta_6 \times \text{Mobile word rate} + \beta_7 \times \text{Time word rate} + \beta_8 \times \text{Degree adverbial rate} + \beta_9 \times \text{Perceptual experience words} + \beta_{10} \times \text{Emotional Experience Words} + \epsilon \]

The specific regression results are shown in Table 2.

| variable                     | Coefficient | Std. error | z     | P>|z| |
|------------------------------|-------------|------------|-------|-----|
| Picture rate                 | 243.6382    | 13.9956    | 17.41 | 0.000 |
| Text Length                  | 0.0000951   | 5.47E-06   | 17.41 | 0.000 |
| Number rate                  | 11.1269     | 0.5351885  | 20.79 | 0.000 |
| Money word rate              | 14.95464    | 0.7407113  | 20.19 | 0.000 |
| Spatial word rate            | 5.166585    | 0.2147477  | 24.06 | 0.000 |
| Mobile word rate             | 1.798729    | 0.4911352  | 3.66  | 0.000 |
| Time word rate               | 4.256614    | 0.5221368  | 8.15  | 0.000 |
| Degree adverbial rate        | 21.12831    | 0.6267988  | 33.71 | 0.000 |
| Perceptual experience word rate | 8.942494 | 0.4049135  | 22.08 | 0.000 |
| Emotional experience word rate | 3.67807   | 0.7580755  | 4.85  | 0.000 |

\( LR \text{ test of } \alpha=0: \text{chibar2}(01) = 2.7e+07 \quad \text{Prob } \geq \text{chibar2} = 0.000 \) (likelihood ratio test)

According to the above table, first of all, the overall validity of the model is analyzed, and the original hypothesis of the model test here is that the quality of the model is the same in both cases, whether or not to put in the independent variables and control variables, which are eleven variables; from the above table, it can be seen that the p-value of the likelihood ratio test is 0 is less than 0.05, thus indicating that the original hypothesis is rejected, i.e., the independent variables put in this construction of the model have validity, and this model construction is meaningful. The final results show that.
Analysis of the Influence of Text Features on the Usefulness of Information: A Case of Tourism Text

The value of the regression coefficient for the picture rate is 243.6382 and statistically significant, indicating that the picture rate has a significant positive effect on the number of collections, and the hypothesis H1 holds. Similarly, the regression coefficient value of text length, number rate, money word rate, spatial word rate, mobile word rate, number of collections, time word rate, perceptual experience word rate, and emotional experience word rate all show statistical significance, indicating that all the hypothesis from H2 to H10 all got supported.

Based on the above regression analysis, it can be further concluded that with the number of readings as the control variable, a total of ten items, including figure rate, text length, number rate, money word rate, spatial word rate, mobile word rate, time word rate, degree adverb rate, perceptual experience word rate, and emotional experience word rate, have a significant positive effect on the number of collections, and all the research hypotheses are valid. The higher the figure rate, number rate, money word rate, spatial word rate, moving word rate, time word rate, degree adverb rate, perceptual experience word rate, and emotional experience word rate, and the longer the text length, the higher the information usefulness. Among them, the figure-containing rate has the greatest influence, and the more influential ones are the emotional experience word rate, money word rate, and degree adverb rate.

5. Conclusion and implications
The results of this paper show that with the number of readings as the control variable, the ten text features of picture rate, text length, number rate, money word rate, spatial word rate, mobile word rate, time word rate, degree adverb rate, perceived experience word rate, and emotional experience word rate all have a significant positive influence relationship on information usefulness. Among them, the greatest degree of influence is found in the picture rate, and the greater degree of influence is found in the emotional experience word rate, money word rate, and degree adverb rate.

The results of this paper suggest the following suggestions for how travel platform or UGC users can create influential and high-quality travel text. (1) The longer the text length of the guide text, the more valuable information it contains. The data shows that the average length of the text of the top 100 tips is 6,674 words, while the average length of the text of the bottom 100 tips is only 2,137 words. (2) Text content should be updated in a timely manner. Tourists who choose free travel have high requirements for the timeliness, accuracy, authenticity and validity of travel tips, so information about new policies, new attractions, new routes, etc. should be updated in a timely manner; timely tips can attract more users to travel and better promote tourism brands and products. (3) Grasp the characteristics of the destination and the needs of the target group, and write travel tips for different destinations and different groups of tourists. Travel tips should be comprehensive and distinctive, so as to better stimulate their desire to travel.

Acknowledgement. The author would like to thank the reviewers for their valuable comments on this paper.

REFERENCES
1. P. Chatterjee, Beyond CPMs and clickthroughs, Understanding consumer interaction with web advertising. Internet Marketing Research: Theory and Practice, (2001) 209-
217.


6. Z.X.Liang, C.X.Liu and Y.Wang, Does the longer and more extreme online reviews make them more useful? Another discussion on the usefulness of online hotel reviews, *Tourism Forum*, 13(03) (2020) 46-57.


