The 6th International Conference on Economic Management and Green Development (ICEMGD 2022)
DOI: 10.54254/2754-1169/3/2022852

Optimization of Consumer Satisfaction Towards Tourism Attractions through Text Mining

Xin Wen[1,a,*], Yijie Li[2,b]

[1]South China Normal University, Guangzhou, China
[2]University of Washington, Main University, Seattle, WA, USA
a. a1552137788@163.com, b. yijie666@uw.edu
**corresponding author*

Abstract: With the ascension of tourism e-commerce, a growing number of consumers are placing greater emphasis on consumer satisfaction and posting their travel experi-ences on tourism e-commerce platform. Processing the online information gener-ated by consumers through text mining has also become a key technology applied in E-commerce. In order to identify the factors that impact consumer satisfaction and find the method to optimize consumer satisfaction, this study collected data from Ctrip.com and applied topic modeling and regressions to online reviews of tourism attractions. The results show that: (1) consumers with higher involve-ment in tourism websites are more likely to have lower satisfaction with tourism attractions. (2) online reviews that focus more on Price, Entrance and tickets tend to have lower satisfaction, while online reviews that contain more about Scenery, History and culture, Entertainment and Service are more likely to be higher on consumer satisfaction. These findings show that text mining can be instrumental for E-commerce to optimize management strategies.

Keywords: Consumer Satisfaction, Topic modeling, Online review, Tourism e-commerce platform.

1. Introduction

In the current era of big data, voluminous user-generated content (UGC) has emerged on tourism e-commerce websites, providing us with a new approach for exploring and analyzing consumer behaviors and social phenomena. UGC contains much more reliable information reflecting numerous valid information that might appeal to potential consumers. Meanwhile, tourism e-commerce platform have proliferated over the past two decades. Among those websites, one of the best-developed tourism e-commerce websites in China is Ctrip.com. It offers a bulk of travel products, booking services as well as travel information and provides a platform for consumers to share their opinions through travel notes, Q&A, online reviews, pictures, likes, and other forms. Therefore, online reviews from tourism e-commerce websites are crucial to the tourism industry because online data allows tourism managers to receive more beneficial feedback to optimize their management. The study of consumer satisfaction can allow managers to provide marketers with a more accurate portrait of their consumers and make it easier to develop effective promotional strategies. This paper aims to determine how consumer experience and consumer characteristics affect consumer satisfaction. We collect data from Ctrip.com from 2002 to 2021. In order to investigate what dimensions of consumer

DOI: 10.54254/2754-1169/3/2022852

experience affect consumer satisfaction, we will use topic modeling to extract several topics from online reviews, namely Price, History and culture, Ticket and entrance, Entertainment, Service, and Scenery. By using topic modeling, the bulk of unstructured data in online reviews can be processed into well-structured and quantitative data. The well-structured data will be applied to OLS regression models, the order probit model and the heterogeneity analysis in order to investigate the correlation between consumer experience, consumer characteristic and consumer satisfaction. We will show how text mining methods can be applied to tourism online reviews and extract the embedded dimensions which can draw managerial insights for tourism attractions' management.

This paper is organized as follows: Section 2 reviews the relevant work. Section 3 explains the data resource and methodology. Section 4 shows the data analysis and results. Conclusions, future work, and limitations are all discussed in Section 5.

2. Literature Review

2.1. Online Consumer Review

Review websites have increased since they change the process of consumer decision-making. The researches on consumer review appear to focus on two aspects: (1) consumer decision-making and (2) product sales [1]. Many studies have highlighted the relationship between consumers' reviews and online sales. Ye et al. found that consumer reviews positively impact online sales, with a 10 percent increase in consumer review ratings boosting online bookings by more than five percent [2]. Research conducted by Chevalier & Mayzlin showed that, by using a difference-in-difference model, book sales are primarily impacted by word-of-mouth on Amazon.com and Barnesandnoble.com [3]. Based on an experiment studied by Vermeulen & Seegers, exposure to online reviews of consumers towards the hotels can increase hotel awareness [4].

The influential factors regarding consumer decision-making have been fixed on well-structured, quantitative factors and unstructured, textual factors. And user-generated content are more voluminous and detailed. Combining with quantitative factors and reviews can constitute a complete research framework to explain the consumer response [5]. For instance, By using the technique of text mining, the determinants that impact consumer satisfaction toward hotels are different and specific to particular types of hotels [6]. Drawing on dual-process theories on how consumers process online reviews of tourism attractions through sentiment analysis, Bigne et al. proposed that review expertise is related to both free and paid-for attractions. [7].

Meanwhile, online reviews seem increasingly significant for tourism attractions that quest to avoid the stagnation phase through allocating resources with priority. UGC may help tourism attractions to grasp the trend by providing up-to-date information on tourism attractions from other consumers [8]. Van de Zee et al. tested the applicability of User-generated content for destination management by analyzing the restaurant reviews from five Flemish art cities and emphasized that the usage of DMOs and BI can improve attractions and allocation [9]. Although numerous studies state the relationship between the online review and consumer behavior, the studies have so far focused more on hotels than consumer attractions in the field of consumer reviews on tourism e-commerce websites.

2.2. Consumer Satisfaction

The concept of consumer satisfaction has been discussed by many scholars in management, economy, marketing, and tourism. Prior research in consumer research has defined individual satisfaction as an assessment of the whole experience of consumption and a cumulative construct that is affected by expectations and performance perceptions [10]. In tourism, consumer satisfaction was defined as a cognitive reflection and real experience towards the tourism service generated by consumers [11]. Consumer satisfaction can be impacted by numerous factors such as amenities, location, transactions.

DOI: 10.54254/2754-1169/3/2022852

Oliver introduced consumer satisfaction as a function of expectation and expectancy disconfirmation [12].

The most traditional approaches used on this topic are face-to-face interviews, mail surveys, questionnaires. In the application of e-commerce, the standard methods to obtain consumer feedback are comments cards, mystery shoppers, and marketing surveys [13]. However, consumers are still limited to explaining their actual and prompt feelings. The cost of producing and issuing questionnaires and incentives is very high. By contrast, the consumers' emotion in online reviews are more abundant and intuitive, allowing consumers managers to directly see the consumers' innermost feelings and prioritize resource allocation of consumer attractions. Zhao et al. demonstrated that a higher level of diversity and sentiment polarity of online hotel reviews result in higher overall consumer satisfaction [14].

2.3. Consumer Factors

Consumer expertise: Many consumer destination managers are engaged in increasing the consumer expertise levels of propagandistic bloggers by changing the way reviews are presented. Reviewer expertise is the perceived diagnosticity of attributes and benefits and the process of conducting product information. Based on the theories from information systems and personality psychology, Wang & Cole investigated how the reviewer's expertise affects reviewer satisfaction and found that a leisure trip positively moderates the impact of reviewer expertise on satisfaction [15]. There are similarities between product satisfaction and consumer satisfaction. However, Yin et al. proposed that online stores are apt to receive a lower rating of the product and more negative content from reviewers with high professional skills [16]. Moreover, more high-expertise reviewers are much more intent to intentionally express negative because they are more capable of processing the intricacies of products [17]. However, there is still controversy on this issue in the academic. Meanwhile, in the study of consumer attractions, social media directly impacts consumer decision-making. Therefore, the managers of consumer attractions must figure out this issue clearly to be conducive to online marketing.

H1a: Consumers' expertise affects consumer satisfaction negatively.

H1b: Consumers' expertise affects consumer satisfaction positively.

Consumer involvement: consumers with high involvement in the online review community are more experienced in traveling to different places and providing a more professional online review. Liu & Park stated that consumers with high involvement in the online review could assess products and services objectively [1]. Potential consumers will seek more useful, objective, professional consumer reviews of tourism attractions from high-involvement travel experts to decide whether they go to the tourism attractions. Furthermore, this motivates more travel bloggers or review experts to post more professional reviews and gain more fans to realize the purpose of raising their network reputation [18]. By analyzing textual reviews from TripAdvisor.com, Zhao et al. studied the role of the reviewer's involvement in consumer satisfaction and found that consumers' review involvement positively impacts their overall satisfaction [14].However, if the research object is a consumer attraction, it is still unclear whether consumer involvement will affect the evaluation of consumer attraction ratings.

H2a: Consumer involvement affects consumer satisfaction negatively.

H2b: Consumer involvement affects consumer satisfaction positively.

2.4. Online Review Factors

Consumer experience of online review: The semantic analysis of online review is widely used by marketing and management. Mankad et al.adapt the methods of topic modeling to process the textual

DOI: 10.54254/2754-1169/3/2022852

hotel online review to investigate the influential factors: amenities, location, transactions, value, and experience [13]. Furthermore, Ahani et al. used the Latent Dirichlet Allocation (LDA) model to explore medical travelers' satisfaction with medical travel [19]. Five topics were found from the online textual review to influence medical travelers' satisfaction, namely outcome of treatment, quality of care, the value of money, patient communication, and hospital/clinic environment. In tourism management, Luo et al. advanced China's 5A global geoparks by incorporating the improved LDA model and the importance-performance analysis (IPA) model through analyzing online textual reviews [20]. The research found that there are ten attributes and 80 elements of attributes, which are history and culture, tour services, well-known degree, travel cost, way of tour, natural scenery, transportation and accommodation, emotional experience, and symbolic features, from 125032 online reviews towards 5A global geoparks. Four negative attributes were specified from the identified attributes, namely travel cost, tour services, well-known degree, and transportation and accommodation [19]. Therefore, the semantic content of online reviews is an essential factor in attracting consumers. From the observation of the online review characteristics in this data, when people talk more about ticket reservations, they usually make online reviews such as "long queue time", "changeable epidemic prevention policies", and "complicated booking procedures." In contrast, the positive reviews express the consumers' feelings about the scenic spot itself. Therefore, we propose the following hypothesis:

H3: Consumer's experience affects consumer satisfaction regarding Price, Entertainment, Entrance and ticket, Scenery, Service, and History and culture.

2.5. Textual Mining in Online Review

Topic modeling: Previous studies mainly analyzed language by the amount of information in the online review or the depth of online review, such as topic modeling analysis and thematic-based analysis [21]. Topic Modeling is a set of algorithms used to extract "hidden" topics or themes from large amounts of text information or a set of documents for generalization [22]. LDA model is widely used in topic modeling algorithms. LDA model uses probabilistic frameworks to extract hidden topic structures from the cleaned text. All online reviews share the same topic set, and each online review is a mixture of different topics [13]. This paper first uses the Chinese thesaurus jieba for word segmentation. Second, then we apply an LDA model in topic modeling. And then, we employed P(topic|review), which represents the probability distribution of topics for a given review as a semantic variable for online review, and P(word|review), which represents the probability distribution of words for a given topic. However, most topic modeling approaches suffer from the imperfection of assuming that the ideal number of topics is a given value, so an external or iterative process is required to determine the optimal number of topics to achieve meaningful results [23]. Zhao & Chen. proposed a perplexity-based method to evaluate the number of topics in a better fit [24].

3. Data Description

3.1. Data Resource and Data Cleaning

We use Python to capture a total of 200058 tourism attractions reviews on Ctrip.com from February 2002 to February 2021. Combined with the geographical location and hot spots ranking of Ctrip.com, we select the hot consumer cities in China. Fifty-seven popular tourism attractions in eight cities were covered, and 200,058 online reviews data were collected. Afterward, we cleaned the dirty data sets and removed the messy code such as emojis, garble, and spaces. Then we divided the sentences into words by analyzing the data with Jieba lexical database and got 190012 online reviews. We dropped the reviews of less than ten words and conducted LDA topic modeling on the textual reviews. Finally, we used a regression model and robust test.

DOI: 10.54254/2754-1169/3/2022852

Table 1: Summary of the dataset.

Item	Value
Name of data resource	Ctrip.com
Number of tourism attractions	57
Total number of consumer review	200058
Total number of cleaned consumer review	190012
Total number of consumer review more than ten words	172,264

3.2. Variable

The main structure of variables in this paper is divided into independent variables, dependent variable and control variables. The dependent variable is the rating of consumers on tourism attractions and ranks from 1 to 5, which represents consumer satisfaction.

Independent variables. The independent variables are divided into three parts, namely consumer experience, consumer expertise and consumer involvement. Consumer experience variable is derived from the analysis of online reviews using Topic Modeling, which shows the percentage of certain topic that be mentioned in online review. A higher proportion means consumers pay more attention to this topic. consumer experience includes six topics in terms of price, entertainment, entrance and ticket, scenery, service, history, and culture, which range from 0 to 1. Consumer expertise is represented by the number of cities a consumer has visited in total, while consumer involvement refers to the activation of a traveler on tourism e-commerce websites and is represented by the cumulative number of likes a consumer has received on tourism e-commerce websites.

Control variables. In this paper, control variables are selected from the factors of content length, number of like, picture, the number of fans and the number of subscriber. Specifically: (1) content length refers to the number of words in a online review. (2) The number of likes refers to the number of likes on each online review. (3) The picture is a dummy variable. The value is 1 if there is a picture in online review, and 0 if there is no picture. (3) The number of fans and the number of subscriber are both characteristics of consumers themselves.

Table 2: Description of the datasett.

Type	Variable	Mean	Std.Dev	Min	Max
Review derivative variables	Topic 1: Price	0.068	0.109	0	0.976
	Topic 2: Entertainment	0.113	0.164	0	0.991
	Topic 3: Entrance and Ticket	0.132	0.175	0	0.988
	Topic 4: Scenery	0.075	0.152	0	0.978
	Topic 5: Service	0.118	0.153	0	0.991
	Topic 6: History and Culture	0.118	0.172	0	0.967
consumer's characteristic	consumer's involvement: Total_like	919.517	7466.987	0	472709
	In (total like)	1.232	2.643	0	13.066
	consumer's expertise: Number of cities	9.104	36.58	0	672
Control variables	Review helpfulness: Number of like	1.562	19.407	0	2850
	Picture (yes,1; no,0)	0.409	0.492	0	1
	Number of fans	16.216	321.623	0	14002
	Number of subscribers	81.212	1678.526	0	143924
Dependent variable	consumer satisfaction	4.2	1.386	0	5

DOI: 10.54254/2754-1169/3/2022852

Table 3: Words in different topics.

Rank	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
	Price	Entertainment	Entrance and Ticket	Scenery	Service	History and Culture
1	Rip off	play	entrance ticket	scenery	dining	heritage
	0.023	0.101	0.022	0.099	0.022	0.038
2	economical	like	ticket	beautiful	service	ancient
	0.017	0.057	0.020	0.084	0.016	0.017
3	worth	fun	Yu Garden	nice	accommodation	architecture
	0.014	0.034	0.016	0.070	0.016	0.014
4	location	entertainment	Ctrip	high	buffet	Scenic spot
	0.013	0.031	0.015	0.066	0.016	0.011
5	worthy	children	queue	overall	hotel	history
	0.011	0.031	0.014	0.053	0.011	0.011
6	price	happy	scenic area	snow	breakfast	street
	0.011	0.021	0.012	0.044	0.010	0.010
7	deserve	suitable	scenic spot	fun	lunch	garden
	0.011	0.018	0.012	0.035	0.010	0.010
8	worse	location	time	weather	beautiful	culture
	0.010	0.016	0.010	0.025	0.009	0.009
9	expensive	kids	in advance	fabulous	lake	feeling
	0.009	0.014	0.010	0.024	0.008	0.009
10	time	next time	park	low	bed	feature
	0.009	0.013	0.009	0.020	0.008	0.008
11	satisfy	interesting	night	night scene	time	city
	0.009	0.012	0.008	0.019	0.008	0.008
12	cheap	guide	weather	interesting	pleasant	millennium
	0.008	0.011	0.008	0.017	0.008	0.007
13	attitude	amuse	free	amusing	summer	year
	0.007	0.011	0.006	0.017	0.007	0.007
14	crowd	tour	travel agency	pretty	arrangement	style
	0.007	0.011	0.006	0.017	0.007	0.006
15	central	driver	buy	rip off	schedule	technology
	0.007	0.010	0.006	0.014	0.007	0.006
16	beautiful	return	purchase	environment	delicious	China
	0.006	0.008	0.006	0.012	0.006	0.006
17	nation	enthusiasm	reserve	air	landmark	grand
	0.006	0.008	0.005	0.009	0.006	0.006
18	traffic	friend	internet	fresh	snack	flourishing
	0.006	0.007	0.005	0.008	0.005	0.005
19	park	knowledge	scan	sky	chef	deserve
	0.006	0.007	0.005	0.006	0.005	0.005
20	bad	attitude	ID card	firework	photograph	wisdom

3.3. Text Analysis Used in Online Consumer Review

Topic Modeling

We employ Gibbs sampling for the perplexity computation and choose the number of the topic of online review prior to gaining the perplexity score, which can measure the goodness of fit of the number of topics. A lower score of "perplexity" means a better fit. As can be seen from the figure 1, the perplexity scores of topics from 4 to 11 are all deficient. We select six topics as the perplexity number by combining the actual meaning of topics and perplexity scores. The LDA topic modeling algorithm outputs posterior probabilities of words for each topic. The probability of each word indicates the likelihood that the word will be assigned to the topic. This article ranks each word

DOI: 10.54254/2754-1169/3/2022852

according to its probability, and we choose the top 20 topic words. We input the entire data set into this model to obtain the probability distributions for each review on six topics.

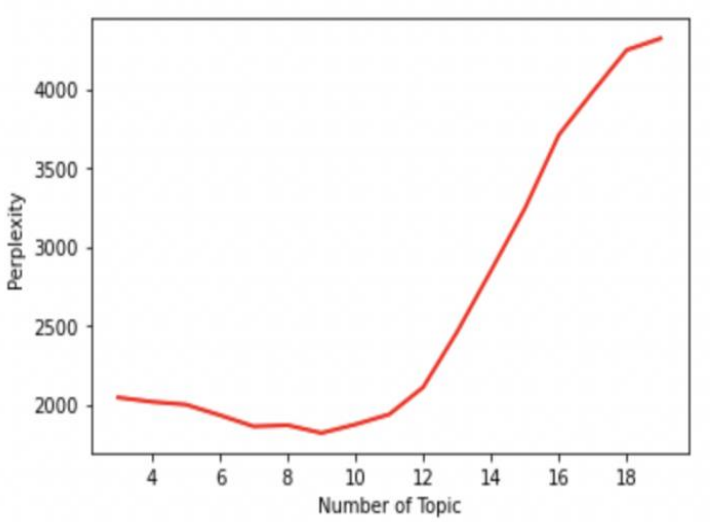

Figure 1: The accuracy of CNN model.

4. Data Analysis

4.1. Research Model

The regression model is:

$$Consumer\_satisfaction = \beta_{1\ -5}\ Consumer\_Experience + \beta_6 Tourist\_expertise + \beta_7\ Consumer\_involvement + \beta_i\ Control\_variables + \varepsilon \quad (1)$$

where Consumer_satisfaction is a response, $\beta_i$are parameters. x_i , (i = 1,.....,n) are variables. ε is a random perturbed variable.

Consumer_experience are the main explaining variables and includes a set of factors, covering Price, Entertainment, Entrance and ticket, Scenery, Service, History and culture. $\beta_{1\ -5}$ represents the influence of six dimensions of consumer experience on consumer satisfaction. Xonsumer_expertise is the number of cities that consumer have visited. Consumer_involvement is the number of likes that consumer have received from tourism e-commerce websites. Control_variables are composed of content length, review helpfulness, number of fans, number of subscribers. The model aims to identify β_1 to β_{10}, the influential factors of consumer satisfaction.

4.2. Data Analysis

We use the OLS model and fix effect to investigate the relationship between consumer experience, consumer characteristics, control variables, and consumer satisfaction. In each regression model, the dependent variable is consumer satisfaction (the rating towards tourism attractions of online review), and we changed the independent variables and the type of models, as shown in Table 4. Models 1 are the independent variables in the hypothesis of this paper, and control variables are added in model 2 and model 3 to improve the r-square interpretation of the whole regression model. Subsequently, in order to control the effect of different tourism attractions, we adopted the time fix effect model in model 4 and region fix effect in model 5. At the same time, we also use the time-fixed effect model to test the reliability of the regression model (model 6). We take logarithms of total_like, and

DOI: 10.54254/2754-1169/3/2022852

content_length to increase the coefficient. In order to test the robustness of the model, we also applied the Ordered Probit model in the data analysis. We'll explain why we use the Ordered Probit in next paragraph. To solve the potential heteroscedasticity problem, we perform logarithmic processing on variables, including the number of like in total, the number of cities, the number of fans and the number of subscriber.

Table 4: Introduction of models.

Model	Independent variables	Model Type
1	Topic1~6, consumer characteristic	OLS
2	Topic1~6, consumer characteristic, Control Variables	OLS
3	Topic1~6, consumer characteristic, Control Variables, ln(Total_like), ln(number_of_city), ln(number_of_fan), ln(number_of_subscriber)	OLS
4	Topic1~6, consumer characteristic, Control Variables, ln(Total_like), ln(number_of_city), ln(number_of_fan), ln(number_of_subscriber)	OLS(time fix effect)
5	Topic1~6, consumer characteristic, Control Variables, ln(Total_like), ln(number_of_city), ln(number_of_fan), ln(number_of_subscriber)	OLS(region fix effect)
6	Topic1~6, consumer characteristic, Control Variables, ln(Total_like), ln(number_of_city), ln(number_of_fan), ln(number_of_subscriber)	OLS(time*region fix effect)
7	Topic1~6, consumer characteristic, Control Variables, ln(Total_like), ln(number_of_city), ln(number_of_fan), ln(number_of_subscriber)	Order Probit (Robustness check)

As shown from table 5, 0.05 was selected as the boundary of significance. Due to the large heteroscedasticity of consumer expertise and consumer involvement, we took logarithms of consumer expertise and consumer involvement. In model 2-7, consumer expertise shows a significant positive correlation with consumer satisfaction, and consumer involvement shows a significant negative correlation with consumer satisfaction. In H2, we found that, as expected, consumer involvement has a negative impact on consumer satisfaction. In H3, we found that when Topic 1 added one standard error in model 4, consumer satisfaction would decrease by 0.0918 standard errors. When topic 3 increases by one standard error in model 4, the satisfaction of consumers will decrease by 0.142 standard errors. We also found that the higher the proportion of Scenery, History and culture, Entertainment, and Service in online reviews, the higher the consumers tend to give higher scores to tourism attractions. The more consumers talk about Entrance and tickets and Price, the more likely they are to give negative comments on tourism attractions. When we observe the original online review, we can find that when the reviews mention the features of tourism attractions, they are mostly high-rating reviews. When the reviews mentioned prices and tickets a lot, the reviewers talked more about the unpleasant experience of traveling rather than focusing on the tourism attractions themselves. Among all OLS models, the topic variable of Service has the greatest impact on consumer satisfaction. In order to test the robustness of the model, we added the order probit model and the results are consistent with OLS model.

Consumer satisfaction is the dependent variable and ordered variable. A score of 1 indicates that the consumer is very dissatisfied, and a score of 5 indicates that the consumer is very satisfied. From 1 to 5, there is a trend of increasing satisfaction. The Ordered Probit model is often used in econometrics to deal with the relationship between ordered variables and is to use observable ordered reflection data to study the law of potential variables that cannot be observed. The Ordered Probit model assumes that the error terms follow the standard normal distribution. The dependent variable in this paper was an ordered discrete variable, and the number of samples was large enough (the Logistic distribution was close to normal distribution). Therefore we employed the Ordered Probit

DOI: 10.54254/2754-1169/3/2022852

Table 5: Summary of the OLS, OLS(with fix effect) and order probit regressions model.

	Model 1	Model 2	Model 2	Model 3	Model 4	Model 5
Model	OLS	OLS	OLS, region fix effect	OLS, time fix effect	OLS, time*region fix effect	Order Probit
consumer experience:						
Topic 1: Price	-0.139***	-0.128***	-0.134***	-0.0412**	-0.0918***	-0.200***
	(0.0177)	(0.0179)	(0.0179)	(0.0177)	(0.0174)	(0.0310)
Topic 2: Entertainment	0.189***	0.185***	0.203***	0.132***	0.172***	0.640***
	(0.0121)	(0.0121)	(0.0121)	(0.0120)	(0.0118)	(0.0242)
Topic 3: Entrance and ticket	-0.217***	-0.215***	-0.220***	-0.188***	-0.142***	-0.265***
	(0.0116)	(0.0116)	(0.0117)	(0.0116)	(0.0114)	(0.0193)
Topic 4: Scenery	0.135***	0.127***	0.131***	0.0318**	0.0331***	0.371***
	(0.0129)	(0.0129)	(0.0129)	(0.0129)	(0.0126)	(0.0257)
Topic 5: Service	0.0722***	0.0680***	0.0512***	0.0445***	0.0457***	0.132***
	(0.0126)	(0.0126)	(0.0127)	(0.0125)	(0.0123)	(0.0228)
Topic 6: History and culture	0.0979***	0.0995***	0.112***	0.116***	0.125***	0.222***
	(0.0117)	(0.0118)	(0.0119)	(0.0117)	(0.0115)	(0.0219)
consumer involvement: Total_like	-3.79e-06***					
	(3.58e-07)					
consumer expertise: Number_of_city	2.58e-06					
	(6.15e-05)					
ln(Total_like)		0.00812***	0.00492**	0.0213***	0.0217***	-0.00399
		(0.00240)	(0.00240)	(0.00240)	(0.00235)	(0.00428)
ln(Number_of_city)		0.0307***	0.0289***	0.0212***	0.0230***	0.0388***
		(0.00390)	(0.00389)	(0.00386)	(0.00377)	(0.00702)
ln(total_like)	0.00103***	0.00111***	0.00115***	0.00158***	0.00201***	0.00191***
	(9.64e-05)	(9.64e-05)	(9.62e-05)	(9.63e-05)	(9.45e-05)	(0.000195)
Picture(yes,1; no,0)	0.136***	0.139***	0.125***	0.167***	0.157***	0.171***
	(0.00398)	(0.00397)	(0.00401)	(0.00404)	(0.00400)	(0.00717)
Number_of_fan	5.40e-05***					
	(6.22e-06)					
Number_of	-7.12e-06***					
	(1.33e-06)					
content_length	0.000867***	0.000842***	0.000851***	0.000823***	0.000808***	0.00137***
	(2.02e-05)	(2.03e-05)	(2.03e-05)	(2.02e-05)	(1.97e-05)	(3.16e-05)
ln(Number_of_fan)		0.0499***	0.0517***	0.0225***	0.0206***	0.113***
		(0.00271)	(0.00272)	(0.00278)	(0.00273)	(0.00451)
ln(Number_of_subscriber)		0.0522***	0.0562***	0.0437***	0.0423***	0.0937***
		(0.00283)	(0.00283)	(0.00281)	(0.00274)	(0.00518)
Constant	1.708***	1.703***	1.727***	1.773***	1.950***	
	(0.00663)	(0.00662)	(0.00667)	(0.00667)	(0.00678)	
Fixed effect (time)	\	\	\	control	\	\
Fixed effect (region)	\	\	control	\	\	\
Fixed effect (time * region)	\	\	\	\	control	\
Observations	172,264	172,264	172,264	172,264	172,254	172,264
R-squared	0.695	0.695	0.697	0.702	0.718	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

model widely used in the literature for estimation. In model 6, every increase of standard error for Topic 1 will increase the dependent variable that makes "consumer satisfaction =1" by 12.3%. In model 10, when topic 1 increases one standard error, the dependent variable of "consumer satisfaction =5" decreases by 9.75%. This reveals that the differences between different levels of satisfaction are not the same.

DOI: 10.54254/2754-1169/3/2022852

Table 6: Summary of the Order Probit model.

	Model 6	Model 7	Model 8	Model 9	Model 10
VARIABLES	Satisfaction=1	Satisfaction=2	Satisfaction=3	Satisfaction=4	Satisfaction=5
Topic 1: Price	0.123*	0.0155	0.0990*	-0.0104	-0.0975***
	(0.0662)	(0.0635)	(0.0595)	(0.0358)	(0.0352)
Topic 2: Entertainment	-0.0297	-0.453***	-0.679***	-0.743***	0.806***
	(0.0459)	(0.0482)	(0.0554)	(0.0279)	(0.0266)
Topic 3: Entrance and ticket	0.448***	0.102***	-0.390***	-0.0795***	-0.150***
	(0.0316)	(0.0328)	(0.0391)	(0.0243)	(0.0240)
Topic 4: Scenery	-0.654***	-0.460***	0.347***	-0.337***	0.314***
	(0.0620)	(0.0539)	(0.0364)	(0.0274)	(0.0259)
Topic 5: Service	-0.143***	-0.0929**	0.181***	-0.135***	0.179***
	(0.0454)	(0.0434)	(0.0418)	(0.0263)	(0.0255)
Topic 6: History and culture	-0.492***	-0.0628	-0.0748*	-0.187***	0.246***
	(0.0604)	(0.0465)	(0.0442)	(0.0241)	(0.0240)
consumer involvement: Total_like	1.02e-06	1.66e-06	-1.33e-06	-2.74e-06***	2.90e-06***
	(1.53e-06)	(1.43e-06)	(1.09e-06)	(9.02e-07)	(8.07e-07)
consumer expertise: Number_of_city	0.00168***	-0.000626	-0.000154	-0.000420**	-0.000248
	(0.000528)	(0.000553)	(0.000274)	(0.000180)	(0.000180)
Constant	0.309***	-0.455***	-1.157***	-1.386***	-1.953***
	(0.0167)	(0.0165)	(0.0178)	(0.0137)	(0.0154)
Observations	171,232	171,232	171,232	171,232	171,232

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7: Summary of Heterogeneity analysis.

	Model 13	Model 14	Model 15	Model 16
VARIABLES	Type=museum	Type=natural attraction	Type=urban landscape	Type=anthropogenic landscape
Topic 1: Price	-0.179**	0.0468*	0.0339	0.147***
	(0.0800)	(0.0268)	(0.0241)	(0.0432)
Topic 2: Entertainment	0.172***	0.118***	0.125***	0.216***
	(0.0383)	(0.0194)	(0.0196)	(0.0232)
Topic 3: Entrance and ticket	0.121***	-0.159***	-0.141***	-0.448***
	(0.0378)	(0.0181)	(0.0169)	(0.0265)
Topic 4: Scenery	0.00442	0.0307	-0.00228	-0.0108
	(0.0469)	(0.0202)	(0.0191)	(0.0259)
Topic 5: Service	0.131***	0.0709***	0.0290	0.163***
	(0.0477)	(0.0192)	(0.0194)	(0.0241)
Topic 6: History and culture	0.338***	0.144***	0.107***	0.156***
	(0.0357)	(0.0206)	(0.0162)	(0.0326)
Consumer involvement: Total_like	1.81e-06**	1.41e-06***	8.07e-07*	2.24e-06**
	(8.02e-07)	(4.62e-07)	(4.74e-07)	(9.74e-07)
Consumer expertise: Number_of_city	0.000165	-0.000129	4.53e-06	0.000105
	(0.000226)	(9.03e-05)	(8.50e-05)	(0.000179)
Constant	1.510***	2.082***	1.726***	1.732***
	(0.0234)	(0.0104)	(0.00992)	(0.0131)
Observations	13,618	59,724	65,848	32,037
R-squared	0.815	0.635	0.726	0.802

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

DOI: 10.54254/2754-1169/3/2022852

In order to explore the heterogeneity of different types of tourism attractions, 57 tourism attractions were divided into museum, natural attraction, urban landscape and anthropogenic landscape. We divided the online reviews of these tourism attractions into four groups for grouped regression. As shown in the table 7, Topic 3 is positively correlated with consumer satisfaction in Model 13, while Topic 3 is positively correlated with dependent variable in other models. As a large number of words in Topic 3 are queuing, ticket and so on, Topic 3 mainly describes the procedures that consumer get into the tourism attractions. While museum attractions are indoors, other types of attractions are outdoors, which makes natural attractions, urban landscape and anthropogenic landscape all negatively correlated with consumer satisfaction.

5. Conclusion

The main conclusions of this paper are as follows: (1) there is a positive correlation between consumers' involvement in tourism websites and their consumer satisfaction. (2) In the dimension of consumer experience, the more attention consumers pay to Price and Ticket and entrance in their online reviews, the lower their consumer satisfaction will be. When they focus more on Scenery, History and culture, Service, and Entertainment in their online reviews, they tend to have higher consumer satisfaction.

Consumers' feedback is a pivotal source of information for continuous improvement in tourism management, but it is challenging to create a centralized description of the consumer experience because the data is sparse and cluttered. This paper overcomes this problem by using the topic modeling and investigating the relationship between consumer experience, consumer characteristics and consumer satisfaction. In addition, the dimensions of tourism experience can allow managers to increase the lure of tourism attractions, while consumers' characteristics provide marketers with a more accurate portrait of their consumers, making it easier for them to develop effective promotional strategies. It is essential for managers to encourage more consumers to participate in expressing their actual feelings about tourism attractions because it can make the topic of online reviews more obvious and focused. Based on these online reviews, managers can invest more resources in a certain dimension of tourism attractions and improve their offerings based on negative feedback according to different dimensions of consumer experience, while can propagandize positive feedback to better their markets. Therefore, digital reform is necessary to enable managers to obtain timely and accurate market information. As the centerpiece of tourism, tourism attractions should speed up their digital reform. The tourism operators must ensure appropriate resources and processes are in place and avoid falling behind in market changes.

Nevertheless, the study has some limitations which deserve to be extended. There are three limitations. First, we cannot generalize our conclusion because we only adopted data from Ctrip.com. Therefore, we can acquire more online reviews from different websites in our future work. Second, because the number of cities visited is provided by consumers on websites, we cannot determine the authenticity of this information, which causes the insignificance of hypothesis 1 in different models. Other methods that acquire the number of cities through the location of consumers' orders and travel notes can also be considered.

References

[1] *Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. Tourism management, 47, 140-151.*

[2] *Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human behavior, 27(2), 634-639.*

DOI: 10.54254/2754-1169/3/2022852

[3] *Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3), 345-354.*

[4] *Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. Tourism management, 30(1), 123-127.*

[5] *Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for consumer reviews. Marketing Science, 35(6), 953-975.*

[6] *Xu, X., & Li, Y. (2016). The antecedents of consumer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. International journal of hospitality management, 55, 57-69.*

[7] *Bigne, E., Ruiz, C., Cuenca, A., Perez, C., & Garcia, A. (2021). What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations. Journal of Destination Marketing & Management, 20, 100570.*

[8] *Bernini, C., & Cagnone, S. (2014). Analysing consumer satisfaction at a mature and multi-product destination. Current Issues in Tourism, 17(1), 1-20.*

[9] *Van der Zee, E., Bertocchi, D., & Vanneste, D. (2020). Distribution of consumers within urban heritage destinations: A hot spot/cold spot analysis of TripAdvisor data as support for destination management. Current Issues in Tourism, 23(2), 175-196.*

[10] *Johnson, M. D., Anderson, E. W., & Fornell, C. (1995). Rational and adaptive performance expectations in a consumer satisfaction framework. Journal of consumer research, 21(4), 695-707.*

[11] *Baker, D. A., & Crompton, J. L. (2000). Quality, satisfaction and behavioral intentions. Annals of tourism research, 27(3), 785-804.*

[12] *Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. Journal of marketing research, 17(4), 460-469.*

[13] *Mankad, S., Han, H. S., Goh, J., & Gavirneni, S. (2016). Understanding online hotel reviews through automated text analysis. Service Science, 8(2), 124-138.*

[14] *Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall consumer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, 111-121.*

[15] *Wang, J., & Cole, C. A. (2016). The effects of age and expertise on product evaluations: does the type of information matter?. Management Science, 62(7), 2039-2053.*

[16] *Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. MIS quarterly, 38(2), 539-560.*

[17] *Guo, B., & Zhou, S. (2016). Understanding the impact of prior reviews on subsequent reviews: The role of rating volume, variance and reviewer characteristics. Electronic Commerce Research and Applications, 20, 147-158.*

[18] *Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. Decision Support Systems, 81, 30-40.*

[19] *Ahani, A., Nilashi, M., Zogaan, W. A., Samad, S., Aljehane, N. O., Alhargan, A., ... & Sanzogni, L. (2021). Evaluating medical travelers' satisfaction through online review analysis. Journal of Hospitality and Tourism Management, 48, 519-537.*

[20] *Luo, Y., He, J., Mou, Y., Wang, J., & Liu, T. (2021). Exploring China's 5A global geoparks through online tourism reviews: A mining model based on machine learning approach. Tourism Management Perspectives, 37, 100769.*

[21] *Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.*

[22] *Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582.*

[23] *Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of online consumer reviews: Readers' objectives and review cues. International Journal of Electronic Commerce, 17(2), 99-126.*

[24] *Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. In BMC bioinformatics , 16(13), 1-10.*