Thesis for the Master of Science

A Text Mining Approach to Understand Travelers' Destination Choices Using Online Travel Reviews

Ji Yeon Hur

Graduate School of Hanyang University

February 2023

Table of Contents

Abstract

Hur, Ji Yeon
Dept. of Business Informatics
Graduate School of
Hanyang University

With the advent of the digital age, understanding online customer behavior has become a major challenge and is of extreme importance for companies and businesses. According to previous research on the matter, high-involvement products, such as travel products, often have a greater influence on customers' decision-making process than low-involvement products. As such, it is important for businesses that sell high-involvement products to pay close attention to online reviews as it is a particularly influential source of information for customers. Over the past decade, numerous studies were conducted in order to identify the factors that influence travelers' satisfaction with their travel destinations. However, these studies mainly relied on traditional methods, such as surveys, to gather and analyze data. This study aims to investigate and verify the destination attractiveness factors that have been identified in previous studies by using online travel reviews as a main source of data. In order to accomplish this, 'Attraction' data was extracted from TripAdvisor up to December 2021. Once the data was collected, several techniques, including LDA, TextRank, XGBoost, and SHAP value, were used to analyze and identified the features that influence customer and the impact they have on their satisfaction. Finally, two statistical techniques, the Chi-square test and Correspondence Analysis (CA), were used to validate the results obtained. This study has significant implications as it presents a text mining approach using various algorithms and statistical techniques, while verifying the results obtained in past studies, which provide tourism-related practitioners and marketers additional insights to help them develop promotion plans for their region as a travel destination. It can help them tailor their strategies to the attributes of their destination and the preferences of their target customers.

Keywords: Online Travel Reviews, customer satisfaction analysis, Chi-square test, CA analysis, text-mining, LDA, TextRank, XGBoost, SHAP value

Chapter1. Introduction

In the 21st century, the international tourism industry contributes to the growth of the global economy by creating jobs and filling the deficit caused by commodity trade. In 2018, the number of international travel arrivals reached 1.4 billion, an increase of 5%, and the value of exports rose to $1.7 trillion. Moreover, the importance of tourism industry management and development is increasing day by day as there are several factors suggesting greater future tourism industry growth such as future technological advancements (digitalization, improved accessibility, etc.), lower travel costs, and a growing middle class in emerging countries (UNWTO, 2019). Therefore, the efforts of tourism practitioners and marketers to attract and retain travelers are becoming increasingly important over time. According to Reisinger et al. (2009), the main challenge for tourism practitioners and marketers is to identify the characteristics of individual visitors and to effectively promote the characteristics of their destination to them. In this sense, many existing tourism-related studies have been conducted from various perspectives, such as cultural characteristics of travel destinations, destination advertising language style, types of travel destinations (hedonic, utilitarian), cultural background of travelers, travel decision-making behaviors, and travelers' desires and preferences (Reisinger et al., 2009; Vinerean, 2013; Byun and Jang, 2015; Guo et al., 2017; Goffi et al., 2018). One of the dominant streams in the study of tourism considers the importance of the cultural context as a characteristic influencing tourism patterns (Moscardo, 2004; Kang and Moscardo, 2006). In other words, to understand the factors that travelers consider when selecting a destination, it is essential to examine at the cultural context and how this context influences their choices. Through the cultural background of travelers and the cultural characteristics of destinations in the global tourism market, researchers can understand the value that travelers perceive as significant when choosing a destination, and the interrelationship between the cultural background of travelers and their value for the destination (Kang and Moscardo, 2006; Reisinger et al., 2009).

In this context, traveler satisfaction studies have also been actively carried out in the tourism industry. Most of the tourism-related literature in the past has asserted a positive relationship between customer satisfaction and destination loyalty, suggesting the importance of analyzing traveler satisfaction (Morgan and Hunt, 1994; Yoon and Uysal, 2005). Customers tend to be satisfied with service when the attributes of the service quality they deem the most important are met or exceeded. At this time, customers provide positive word of mouth, revisit travel destinations, or recommend travel destinations to acquaintances (Morgan and Hunt, 1994; Yoon and Uysal, 2005; Guo et al., 2017). On the other hand, if travelers are dissatisfied, they will damage the reputation and image of the travel destination by spreading negative word of mouth or considering whether to revisit the destination or not. (Kim et al., 2021). Such negative feedback has a far greater impact on the travel destination selection stage than positive feedback, therefore managing negative feedback is treated as a matter (Kim et al., 2021). By analyzing traveler satisfaction, companies and governments involved in the tourism industry can check whether the expectations of travelers are being met. Through this, it is possible to identify areas for improvement or weaknesses and strengths compared to other competing destinations, enabling more efficient management of travel destinations and attracting travelers (Grigoroudis and Sisikos, 2009).

Past traveler satisfaction studies have applied several methodologies. Typically, conventional research has been conducted using traditional methods such as surveys and interviews to identify and assess traveler's experiences, satisfaction, and expectations. However, this method has a drawback as it is time-consuming and costly for researchers to collect samples and conduct the analysis, as well as it is possible to draw biased conclusions due to the quality of respondents' responses, limited sample size, and inconsistent survey questionnaire items (Guo et al., 2017; Aman, 2021). In order to address these difficulties, recent studies have shown many attempts to uncover traveler satisfaction through user-generated content (UGC) (Leon, 2019; Nakayama and Wan, 2019; Jia, 2020; Hlee et al., 2021). Recent technological developments such as the wide spread of Web 2.0 have made it possible for customers who have purchased products online to directly write about their experiences and quickly share their opinions with future customers. This enables businesses and governments to gain a better understanding to manage related products and services (Grigoroudis and Sisikos, 2009). In addition, these UGC data have the advantage of being easily accessible for free or at a low cost (Guo et al., 2017; Aman, 2021). In particular, the importance of Online Travel Reviews (OTR) is increasing in tourism products that are classified as high customer engagement products. Accordingly, customer satisfaction analysis using OTR has been actively conducted in the recent few years (Gretzel, 2022).

This study attempts to find out how destination attractiveness attributes, which are considered important for travelers to choose a destination, differ according to cultural context and type of destination attraction through online travel reviews. This study aims to answer the following research questions.

RQ1. What are the key features mentioned in online travel reviews?
RQ2. Does the frequency of mention of key features across different cultures (East Asian vs. Western) differ by destination attraction type (Hedonic vs. Utilitarian)?
RQ3. Is there a relationship between features and contextual characteristics of travelers (East Asian / Hedonic, East Asian/ Utilitarian, Western/ Hedonic, Western/ Utilitarian)?
RQ4. What features influence traveler satisfaction significantly when selecting a travel destination based on destination attraction type (Hedonic vs. Utilitarian) and cultural background (East Asian vs. Western)? What differences do they make?

This study makes several contributions to the understanding of the factors that influence travelers. By using user-generated content (UGC) data, such as online reviews, the study is able to gather feedback directly from customers and, therefore, the wisdom of the crowd. Thereby, it was possible to obtain objective results that may be more reliable than those obtained through traditional methods such as surveys and interviews. The study also investigates the influence of cultural background and destination attractions on customer satisfaction and uses statistical methods to verify the results of previous studies on these factors. Finally, this study suggests various text mining approaches including LDA, TextRank, XGBoost, TreeSHAP that can be used by tourism-related organizations. It is hoped that the results of this study will provide tourism-related practitioners with useful insights for creating more effective and detailed advertising plans.

This study is structured as follows: Chapter 2 presents a literature review and the development of hypotheses, highlighting the importance of online travel reviews (OTR) and demonstrating that the characteristics of a reviewer can influence the characteristics of readers (potential customers). It also discusses previous research indicating that cultural characteristics of reviewers can impact customer behavior, as well as the differences in destination selection between East Asian and Western cultures and by attraction type, a hypothesis was then developed and presented. In Chapter 3, various text-mining approaches are introduced, and the results of these analyses are presented in Chapter 4. Chapter 5 interprets the results from Chapter 4, and Chapter 6 offers a summary and conclusion. Finally, Chapter 7 discusses the limitations of this study

Chapter4. Result

By applying the methodology presented above, this study was able to discover the key features of travel reviews and conduct an in-depth analysis of them. In this section, the topic modeling results were verified by mapping with existing studies related to destination attractiveness factors. Moreover, each feature as well as each feature was manually cross-checked by three researchers. Afterward, a chi-square test and correspondence analysis (CA) were set up to examine the significant differences and the relationship between the destination attractiveness factor and the contextual characteristics of the traveler (the attraction type, i.e., hedonic vs. utilitarian, as well as the cultural background, i.e., East Asian vs. Western) (Kim et al., 2021). To perform both analyses, the matrix (reviewer id by features) was generated as a binary value (0,1) depending on whether each feature unique reviewer mentioned it (if it is mentioned it will be '1' otherwise, '0'). In order to conduct the statistical analysis above, SPSS version 27 was used. Afterward, by predicting the reviewer's star rating using the feature proportions, the reviewer's satisfaction index can be determined. This study used the XGBoost with TreeSHAP value to understand which destination attractiveness factors affect travelers' positive or negative emotions.

Before discussing the topic modeling result, Table 2 summarizes previous studies related to destination attractiveness factors to justify and verify the topic modeling result. These factors below were considered when researchers judged the label of features (the result of topic modeling) independently (Debortoli, 2016).

Factor	Description	Reference
Packages	pre-arranged packages by intermediaries and principals	(Cooper et al., 1998; Buhalis, 2000)
Accessibility	entire transportation system comprising of routes, terminals, and vehicles	(Ritchie and Zins, 1978; Cooper et al., 1998; Buhalis, 2000; Reitsamer et al., 2016)
Facilities	accommodation and catering facilities, retailing, other tourist services	(Cooper et al., 1998; Buhalis, 2000; Reitsamer et al., 2016)
Activities	all activities available at the destination and what consumers will do during their visit	(Cooper et al., 1998; Buhalis, 2000)
Scenery	the natural form and landscape of a destination	(Ritchie and Zins, 1978; Milman and Abraham, 1995; Reitsamer et al., 2016)
Climate	amount of sunshine, temperature, winds, precipitation	(Ritchie and Zins, 1978; Hu et al, 1993; Milman and Abraham, 1995)
Safety	level of personal and material safety	(Ritchie and Zins, 1978; Milman and Abraham, 1995)
Local community/ Attitudes towards tourists	natural and experiential resources in a destination and refers to a connection with local people & the warmth of reception by the local population, ease of communication, willingness to provide information, and a lack of hostility towards tourism activities.	(Ritchie and Zins, 1978; Kim et al., 2012; Reitsamer et al., 2016)
Price level	the value received for money spent on major services, food, lodging and transportation within the region	(Ritchie and Zins, 1978; Milman and Abraham, 1995)
Season	seasonal preference of a destination	(Hong-bumm, 1998)
Family oriented	suitability for families with children	(Milman and Abraham, 1995; Hong-bumm, 1998)
Information/ Credibility	destination-related details and images (a factor that gives credibility)	(Hong-bumm, 1998)
Festivals/ Special events	festival and special event in a destination	(Shafiee et al., 2021)

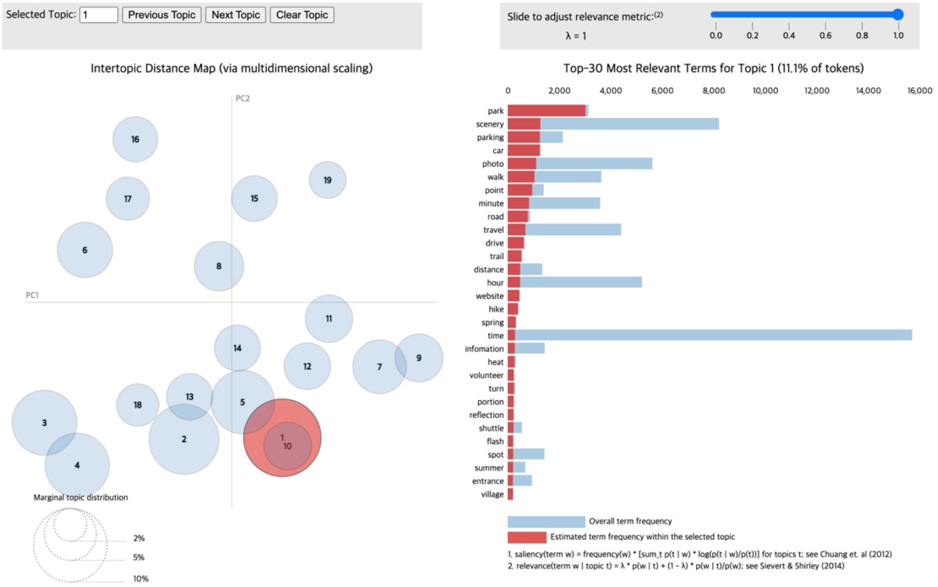
<Table 3> Destination attractiveness factors in prior studies

Section 4.1 Identification of the key features in Online Travel Reviews (OTR)

An example of the extracted keyword process is given in Figure 4 and Table 3. It shows the differences in the top 10 keywords between the three algorithms. When applying the relevance score to the basic LDA model, it can be seen that the top 10 keywords appear to be more focused on words related to 'park', 'car', and 'hiking' than the basic LDA model. Subsequently, in this study, the reviewed documents are separated at the sentence level to assign reviews to each feature. Thus, sentences containing the top 10 keywords to which the relevance score was applied were assigned to each feature. Then TextRank algorithm was applied to them, in regard to topic 1, it was more focused on the '(parking, lot)' and '(car, park)' than the keywords applied to the relevance score. Looking at the keyword selection process, we can see that the gray part in the TextRank keyword column in Figure 5 did not overlap with other topics, so it was deemed appropriate to represent topic 1, so it is confirmed as a topic keyword 1. On the other hand, keywords written in red were keywords that overlapped with other topics and were ignored because they had a relatively lower weight than other topics. In the end, the sentences corresponding to each feature were reassigned using the keywords taken from TextRank.

According to previous studies, in order to label features, the relevant keywords extracted for each feature should be combined with related previous studies to verify them. (Debortoli et al., 2016). In the case of topic modeling algorithms, it is unsupervised learning, therefore human engagement is required to label the features (Ibrahim et al., 2019; Bastani et al., 2019; Aman et al., 2021). Moreover, to select a reasonable label for features, the three researchers independently verified the sentences and keywords selected from TextRank. Finally, the feature was only labeled when all researchers reached agreement on all features (Debortoli et al., 2016; Silge and Robinson, 2017). Simultaneously, the label was considered and compared to previous studies related to the destination attractiveness factors summarized in Table 2 for the final label decision. The 'Y' in the prior study column of Table 4 means that there are prior studies, and the '/' means that there are barely prior studies.

Finally, in order to prevent the appearance of ambiguous features at this stage, features that have a similar meaning (or the sub-content) were clarified by combining them together as a particular feature. Otherwise, it separated (Debutoli et al., 2016; Aman et al., 2021). In this study, topics 10 and 13, which are judged to share the same content, were combined into the 'Activity' feature, and topics 12 and 16 were combined into the 'Scenery' feature to clarify the contents (topics 14 and 18 were eliminated as they were judged to be meaningless). In conclusion, the total number of agreed-upon features is 13. The feature label and review content examples of each feature are summarized in Table 4.


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
16
17
15
19
6
8
11
14
12
7
9
18
13
5
3
2
1
10
4
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Top-30 Most Relevant Terms for Topic 1 (11.1% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
park
scenery
parking
car
photo
walk
point
minute
road
travel
drive
trail
distance
hour
website
hike
spring
time
infomation
heat
volunteer
turn
portion
reflection
shuttle
flash
spot
summer
entrance
village
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

<Figure 5> Visualization of an LDA model with 18 topics using pyLDAvis

Topic 1				
Rank	LDA keywords	Relevance keywords	TextRank keywords	
			words	weight
1	park	park	(parking, lot)	0.22893
2	scenery	car	(car, park)	0.12674
3	parking	parking	(walk,)	0.12236
4	car	point	(point,)	0.07481
5	photo	road	(scenery,)	0.06319
6	walk	drive	(road,)	0.06307
7	point	trail	(time,)	0.05977
8	minute	website	(beach,)	0.05408
9	road	walk	(minute,)	0.05317
10	travel	hike	(drive,)	0.05234

<Table 4> An example of identifying key features (topic 1)

TextRank keywords	Example	Label	prior study
(parking, lot), (car, park), (walk,), (road,)	*...15 minutes or more from the parking lot...* *...Walk 10 to 15 minutes from the parking lot (sneakers are desirable because they walk in the sand)...*	Parking	/
(tour,), (guide,), (site,), (lot, information), (group,)	*...There are tours and bus tours going by plane, but bus tours are tough if they cannot afford time...* *...Well worth the tour and very well organised...* *...They have guided tours in english...* *...Purchase the tour tickets, then obtain tour group time & explorer around before the meeting time...*	Packages	Y
(bus,), (city,), (ride,) (shuttle,), (stop,), (access,)	*...The free bus is always moving, so you can easily move to a superb view site...* *...As the hotel was close from the starting bus stop, we were able to ride a lot, but this bus route is popular and can*	Accessibility	Y

	be cared immediately, so it could no longer get ride from the bus stop in the center of Waikiki...		
(food, drink), (price,), (service,), (bike,), (town,), (quality,), (ticket,), (minute,), (money,), (price, admission), (advance,)	*...Rental services equipments like snorkels $7-15 and lockers $7 are available...* *...Prices are also high...* *...Food is expensive and if you're not camping think a Cabin cost more than...* *...The price was also reasonable...*	Price Level	Y
(morning,), (night,), (sunset,), (time,), (sun,), (weather,), (evening,), (lot, people)	*...Weather was fine and sunshine was very strong...* *...I saw the sunset, but the weather was not good and I could not say the best, but it was still very beautiful...* *...If you are bad at the weather, please be careful about clothing etc...*	Weather	Y
(summer,), (time, period), (ticket, month), (architecture,), (winter,), (design,), (phone,), (spring,)	*...There were many people because it was a summer vacation period, but I should go absolutely...* *...Summer is super busy...* *...I went twice, but it is a good place to go many times, but it is painful for summer...*	Tourist Season	Y
(hotel,), (shopping,), (distance,), (restaurant,), (street,), (shop,)	*...There are also supermarkets that can be accommodated and shopping...* *...Second stop: Looking for a restaurant to eat, although the road (in the wrong road) came to the parking lot of the town, there is a wooden house, there is a network, but in the end, it is still on the restaurant, and eats a Chinese meal...*	Facilities	Y
(security, check), (story,), (camera,), (hand,), (bag,), (event,), (luggage,)	*...It costs admission fee, but it is safe to put the luggage and enter the sea alone...* *...The valuables can be deposited in the locker with a paid, so I was able to swim with confidence even if I was an*	Safety/Security	Y

	individual...		
(lot, fish), (snorkel,), (watch, video), (swim,), (water,), (place,), (purchase, ticket), (variety,), (location,), (lot, activity), (variety, fish), (image,), (age), (interest,), (staff, member)	*...I can swim with tropical fish and you can swim with tropical fish without going to a deep place...* *...You can perform a snorkeling and other water activities!...* *...Learn about activities and entertainment in advance, and you will have a good time...* *...I love outdoor activity, so I strongly recommend everyone to stop by the Red Rock Canyon...* *...This is a bucket list activity...*	Activities	Y
(photo,), (point,), (tour, guide), (drive,)	*...I was crazy and taking a lot of photos...* *...We stopped at all the roadside parking lots to enjoy the scenery and take photos...* *...Great art and beautiful place our favorite part was finding the artwork featured in "Ferris Bueller's Day Off" and taking a few photos...*	Camera (Taking pictures)	Y
(scenery,), (travel,), (foot,), (history,)	*...The scenery spread in front of you is too magnificent, and it is as much as you lose words...* *...However it was an interesting historical tour and we enjoyed the scenery very much, which is beautiful...* *...The courtyard is serene and many people sat along the outskirts as they took a rest and appreciated the scenery...*	Scenery	Y
(family,), (time, year), (child,), (number, people), (adult,)	*...There are many Junior Ranger Programs that children can learn about the Earth...* *...The whole family went together...* *...Thanks to the parasol, the scenery and family photo shoots were also enjoyable...* *...I personally, would never bring young children...*	Family Oriented	Y
(tourist,), (spot,), (movie,), (crowd)	*...Stunning you've seen it in the movies, so you already know what to expect, and yet,*	Media	Y

	it is absolutely stunning to watch... *...This is one of those places that you've seen many times in pictures, on tv, on the internet, but you still go WOW when you see it with your own eyes...*		

<Table 5> Result of the features labeling

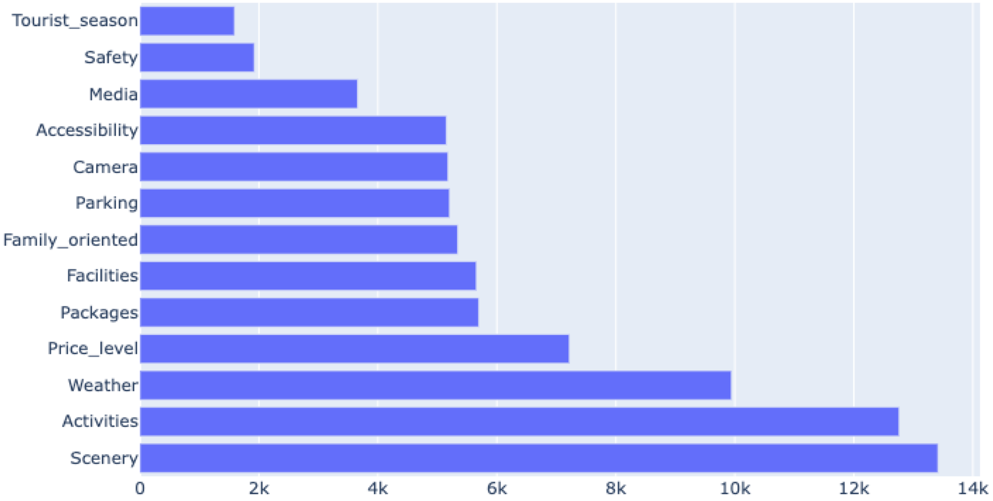
Section 4.2. Distribution of the destination attractiveness factors by contextual characteristics of the traveler

In this study, we assigned each sentence to each feature. Reviews are merged by reviewer ID, which is a unique value, to compare the distribution of mention sizes. Looking into the reviews, several reviewers mention the same feature multiple times. In this case, the reference was counted as 1. That is, if there is more than 1 review assigned to a feature for a reviewer ID, it will be deleted. Therefore, Multi-label text data was created for each reviewer.

When it comes to overall feature distribution, the travel reviewer mentioned "Scenery", "Activities", and "Weather" the most, while "Tourist Season", "Security", and "Media" were mentioned the least. 'Price level', 'Packages', 'Facilities', 'Family oriented', 'Parking', 'Camera', and 'Accessibility' have a medium mention level (Figure 6).

To compare and analyze the difference in the volume of mention according to the contextual characteristics of the traveler, the entire dataset was divided into 4 clusters (East Asian/Hedonic, East Asian/Utilitarian, Western/Hedonic, Western/Utilitarian). A minmax scaler was then applied to the volume to change it to a value between 0 and 1(Figure 7).

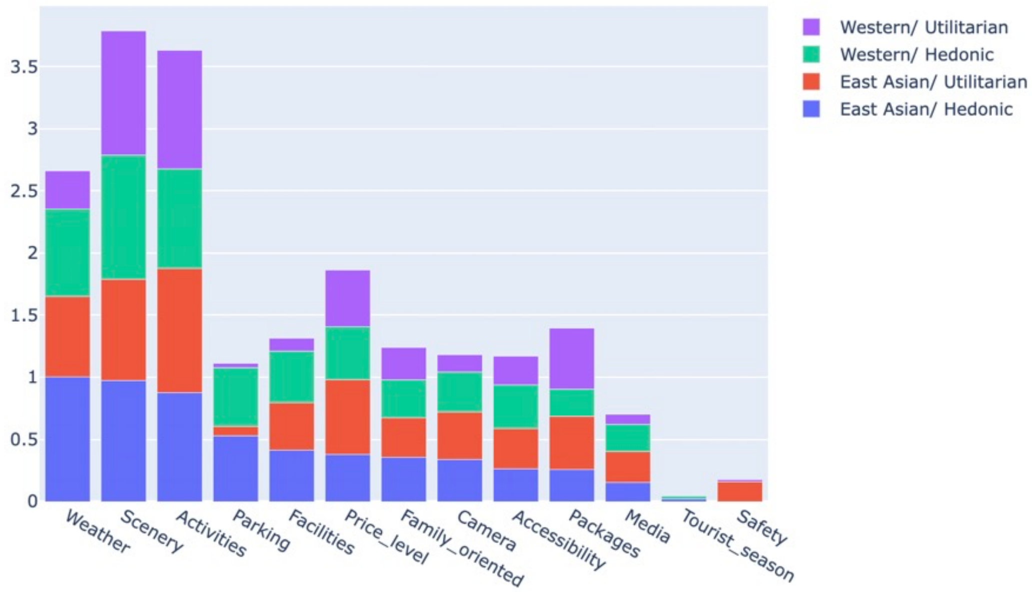
First of all, looking at the most mentioned features ('Scenery', 'Activities', 'Weather'), in the case of 'Scenery', Western has a higher rate of mention of both places (Hedonic vs. Utilitarian) than the East Asian. Furthermore, from the perspective of the type of attraction (Hedonic vs. Utilitarian), it shows a higher rate in the Hedonic place than in the Utilitarian place for both groups (East Asian vs. Western). On the other hand, in the case of 'Activities', the mention rate in East Asia is relatively high in both places, and at the same time, in the Utilitarian place, the rate of mention is high in both groups. Finally, in the case of 'Weather', East Asian is relatively high across all locations, and both groups show a high mention rate in the hedonic location.



<Figure 6> Distribution of overall features

Looking at the medium level of mentioned features ('Price level', 'Packages', 'Facilities', 'Family oriented', 'Parking', 'Camera', 'Accessibility'), in the case of 'Price level', the rate of mention in Utilitarian place is relatively high for both groups. At the same time, in the reviewer's cultural background perspective, the East Asian, is mentioned higher in the Utilitarian place, on the contrary, it is mentioned higher in the Hedonic place for the Western. Next, in the case of 'Packages', the rate of both groups was high in the Utilitarian place, on the other hand, the Western was higher in the Utilitarian place than East Asian and the East Asian was higher in the hedonic place than Western. With regard to 'Facilities' was high in both groups in the Hedonic place, however, in Utilitarian places, the rate of Westerns was remarkably low than East Asian. In the case of 'Family oriented' and 'Parking' both groups appeared high in the Hedonic place, and although there is a small gap of differences between the East Asian and Western, East Asian appears relatively high in all places.

Finally, looking at the least cited features ('Tourist season', 'Safety', 'Media'), in the case of 'Tourist season', both groups have a high rate of mentions in the Hedonic place. On the other hand, 'Safety' appears higher in the utilitarian place for both groups, and the East Asian is remarkably higher than the Western in the utilitarian place. In the case of 'Media', Western is higher in the Hedonic place than East Asian, while East Asian is higher in the Utilitarian place than Western.



<Figure 7> Features distribution by culture & attraction type

Section4.3. Relationships between destination attractiveness factors and contextual characteristics of the traveler

Table 6 is the chi-square test result. In terms of differences by culture, all features except for 'Family oriented' and 'Media' were found to have significant differences, and for Attraction type, all features except for

'Media' were proved to be significant.

Additionally, this study attempted to uncover the relationships between destination attractiveness factors and contextual characteristics of the traveler through CA analysis. To perform the CA analysis, a new table was created with three columns utilizing the table with binary values mentioned above. The three columns are: 1) a column with 13 feature categories ('Scenery', 'Activities', 'Weather', 'Tourist season', 'Safety', 'Media', 'Price level', 'Packages', 'Facilities', 'Family oriented', ' Parking', 'Camera', 'Accessibility'), 2) a column with 4 contextual characteristics categories (East Asian / Hedonic, East Asian/ Utilitarian, Western/ Hedonic, Western/ Utilitarian), and 3) a column with the frequency of each feature. Afterward, the feature category column was changed to multiple choice using the SPSS tool in the table (creating a contingency table between destination attractiveness factors and contextual characteristics of travelers), and a correspondence analysis was performed (Iacobucci and Grisaffe, 2018). The advantage of correspondence analysis is that it can calculate distances and analyze correlations between various categorical variables such as nominal and ordinal scales without assuming the data distribution. Usually, it is preferred to show the relationship between variables in a two-dimensional space, because, in the case of a two-dimensional space, the corresponding relationship between rows and columns can be recognized easily and accurately. When the chi-square statistic is significant, the coordinate points representing the rows and columns in the correspondence analysis result not only move away from each other, but also lie on the opposite of each other on some basis (Ho and Hung, 2008).

In this study, as shown in Table 7, the first dimension and the second dimension explained 78.5% and 15.9%, respectively, and the cumulative explanatory power was 94.4% in total, which was judged to be a sufficient explanatory power. Usually, if the cumulative explanatory power is greater than 70%, it is considered to have explanatory power (Hair et al., 1998). Table 8 shows the explanatory power of the two dimensions, and if the variance of each dimension is less than 50%, it is removed from the CA map (joint plot) shown in Figure 7 (Hair et al., 1998). Therefore, in this study, three features ('Family oriented', 'Media', and 'Accessibility') were removed. For dimension 1, Asian/Hedonic (0.9) and Western/Utilitarian (0.838) explain a large proportion of the variance in the contextual characteristics of traveler, whereas, for dimension 2, Asian/Utilitarian (0.435) and Western/Utilitarian (0.16) explain the relatively large proportion.

In Figure 7, it can be seen that while East Asian and the Western display a similar context in the Hedonistic place, there is a clear difference between East Asia and the West in the utilitarian place. First of all, in the case of the Hedonic place, both East Asian and Western are highly related to 'Camera', 'Facilities', 'Weather', 'Parking', and 'Scenery'. On the other hand, in the case of the Utilitarian place, East Asian is highly related to 'Safety', 'Price level', and 'Camera', and Western is highly related to 'Packages', 'Activities', 'Tourist season', 'Scenery'.

Features	Culture			Attraction type		
	x^2	df	p $value$	x^2	df	p $value$
Parking	4.669	1	.031*	1798.331	1	0**
Packages	46.817	1	0**	413.053	1	0**
Accessibility	50.111	1	0**	6.366	1	.012*
Price level	11.159	1	.001**	77.420	1	0**
Weather	274.904	1	0**	1054.777	1	0**
Tourist season	42.589	1	0**	8.635	1	.004**
Facilities	21.872	1	0**	277.829	1	0**
Safety/Security	16.571	1	0**	316.817	1	0**
Activities	22.219	1	0**	19.709	1	0**
Camera	23.64	1	0**	48.931	1	0**
Scenery	292.945	1	0**	145.156	1	0**
Family oriented	1.1811	1	.181	13.713	1	0**
Media	1.826	1	.178	0.009	1	.929

$*p<.05, **p<.01$

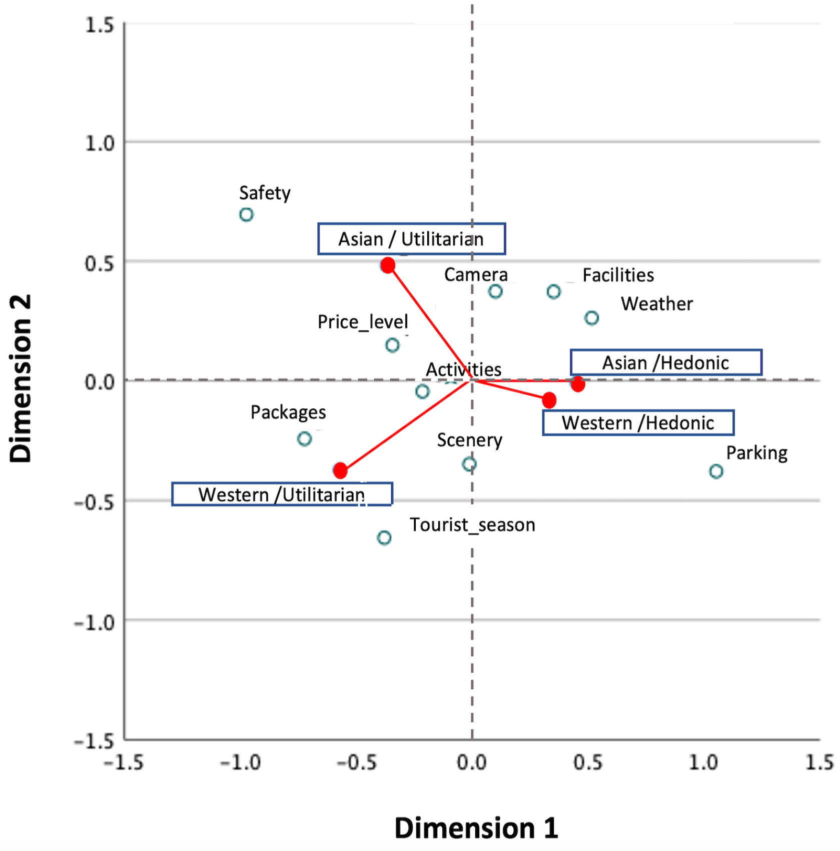
<Table 6> Result of Chi-square Test

			Proportion of Inertia			
Dimension	Singular value	inertia	Explained	Cumulative	Chi-square	Significant
1	0.19	0.036	0.785	0.785		
2	0.086	0.007	0.159	0.944		
3	0.051	0.003	0.056	1		
total		0.046	1	1	3824.266	<.000

<Table 7> Summary table of the CA dimension results

Features	Mass	Dimension 1	Dimension 2	Total
Activities	0.154	**0.948**	0.018	0.966
Facilities	0.068	**0.62**	0.314	0.934
Camera	0.063	0.136	**0.85**	0.986
Family oriented	0.065	0.157	0.307	0.464
Media	0.044	0.151	0.423	0.574
Packages	0.069	**0.917**	0.046	0.963
Parking	0.063	**0.944**	0.055	0.999
Price level	0.087	**0.879**	0.075	0.954
Accessibility	0.062	0.164	0.007	0.171
Safety	0.023	**0.805**	0.185	0.99
Scenery	0.162	0.003	**0.951**	0.954
Tourist season	0.019	0.415	**0.56**	0.976
Weather	0.12	**0.794**	0.093	0.887
Total sum	1.000			
Contextual characteristics	Mass	Dimension 1	Dimension 2	Total
Asian/Hedonic	0.268	**0.9**	0	0.9
Asian/ Utilitarian	0.226	**0.565**	0.435	1
Western/ Hedonic	0.278	**0.791**	0.02	0.811
Western/ Utilitarian	0.227	**0.838**	0.16	0.998
Total sum	1.000			

<Table 8> Contribution table of the CA dimensions



<Figure 8> CA map

Section4.4. Destination attractiveness factors influencing traveler satisfaction

Section4.4.1 Model validity

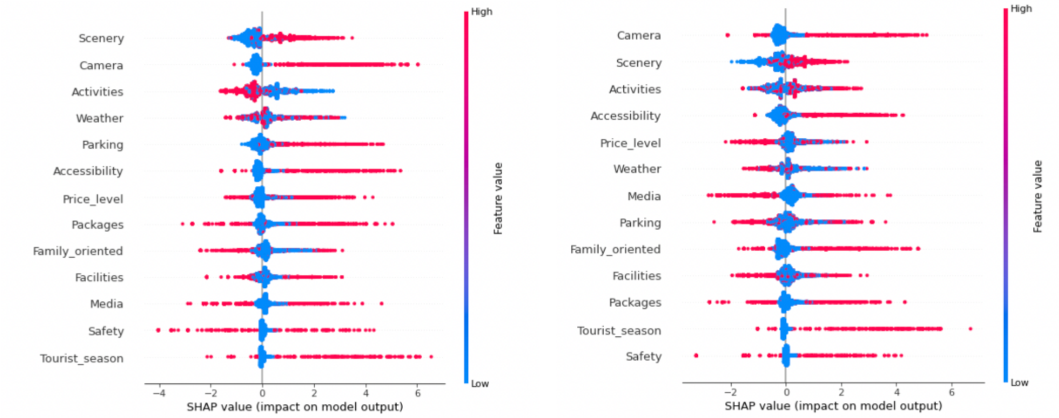
XGBoost was employed in this study to analyze the influence of destination attractiveness factors on reviewers' satisfaction. Before proceeding with the analysis in earnest, it is necessary to balance the size of the target variable to avoid a bias towards a target value (Othman et al., 2018). In this study, an over-sampling method was performed to prevent bias during model training that could occur due to the presence of large differences between the size of the target variables (positive vs. negative). Afterward, overall samples were divided into train/test sets with a ratio of 80/20%, and again the 80% train set was divided into 80/20% as a train set and a validation set, respectively. The hyper-parameters of the model were verified through the validation set, and the optimal model was evaluated by applying the test set. The data set was separated into four clusters (East Asian / Hedonic, East Asian / Utilitarian, Western / Hedonic, and Western / Utilitarian), and each was applied to the analysis separately. After over-sampling, the sizes of samples are 4069, 3828, 4831, and 4870 for each data set. Grid search was applied for hyper-parameter adjustment. The result is shown below.

(1) East Asian/Hedonic dataset: learning_rate 0.35, n_estimaors 250, and max_depth 5 were applied, and evaluation score are f1 78.68%, precision 86.41%, recall 72.22%, accuracy 80.53%. (2) East Asian/ Utilitarian: learning_rate 0.15, n_estimaors 250, max_depth 5 were applied, and evaluation score are f1 80.22%, precision 84.12%, recall 76.66%, accuracy 81.07%. (3) Western/ Hedonic: learning_rate 0.4, n_estimaors 200, max_depth 5 were applied, and evaluation score are f1 80.73%, precision 89.85%, recall 73.29%, accuracy 80.73%. (4) Western/ Utilitarian: learning_rate 0.55, n_estimaors 250, max_depth 5 were applied, and evaluation score are f1 85.09%, precision 84.20%, recall 86.01%, accuracy 85.09%.

Section4.4.2 Comparison of contextual characteristics of travelers through satisfaction prediction

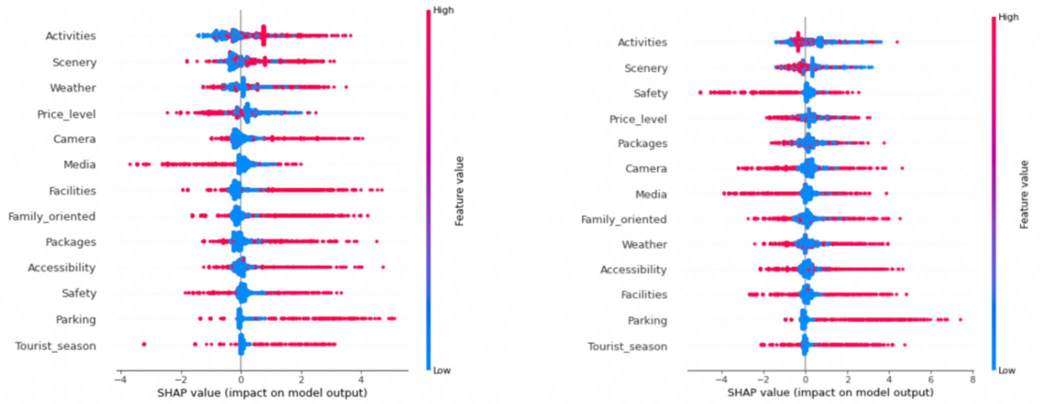
In this section, the influence of destination attractiveness factors on reviewers' satisfaction for each cluster (East Asian/Hedonic, East Asian/Utilitarian, Western/Hedonic, Western/Utilitarian) was confirmed. Through this, it was possible to go beyond a simple frequency analysis and derive the main factors that affect the reviewer's satisfaction for each cluster. Figure 9 and Figure 10 are the results of applying XGBoost to TreeSHAP values. In the graph below, the order of importance of features is ranked in descending order, and the higher the original value of each feature, the redder it is. Therefore, when looking at the results of this study, samples that have a high distribution will appear redder. The horizontal axis of the graph indicates the degree of positive or negative impact on satisfaction, and the more right it is, the more positive it is. The dots that appear here represent each sample (Kim et al., 2021).

In the case of the Hedonic place (Figure 9), 'Scenery', 'Camera', and 'Activities' have a high impact on the satisfaction of East Asian and Western reviewers. The 'Scenery' and 'Camera', both Asians and Westerners show a positive correlation with their satisfaction, On the other hand, 'Activities' negatively correlated with East Asian satisfaction, while it is positively correlated with Western. Next, it was discovered that 'Tourist season', 'Safety', and 'Facilities' have a low impact on the satisfaction of both groups. 'Tourist season" is positively correlated with the satisfaction of both groups, while 'Safety' was relatively negatively correlated with East Asians than with Westerners. As for 'Facilities', the interpretation of the results seems ambiguous. Then, it turns out that 'Parking', 'Packages', 'Media', and 'Accessibility' have different effects on satisfaction on East Asians' and Westerners' satisfaction. Features that affect satisfaction for East Asians rather than Westerners are 'Parking' and 'Package'. "Parking" is positively correlated with East Asians, and on the contrary, it has a negative correlation with Westerners. As for 'Packages', it is ambiguous to interpret. Lastly, in the case of 'Media' and 'Accessibility', it had a relatively higher influence on the satisfaction of Westerners than East Asians. The two features seem ambiguous to interpret.



<Figure 9> SHAP feature contribution in the Hedonic place (East Asian (left) vs. Western(right))

In the case of the Utilitarian place (Figure 10), 'Activities', 'Scenery', and 'Price level' have a high impact on the satisfaction of East Asian and Western reviewer. The 'Activities' shows a positive correlation with East Asians while it shows a negative correlation with Westerners. 'Scenery' is negatively correlated with Westerners while it is positively correlated with East Asians. In terms of 'Price level', it appears to be negatively correlated with East Asian. Next, it was discovered that 'Parking' and 'Tourist season' have a low impact on the satisfaction of both groups. Then, it is found that 'Weather' and 'Facilities' have different effects on the satisfaction of East Asians and Westerners. 'Weather' and 'Facilities' are the features that have influenced East Asians more than they have influenced Westerners. In the case of "Weather" and "Facilities", it shows a positive correlation with both East Asians and Westerners. Lastly, 'Safety' and 'Packages' have a relatively significant impact on the satisfaction of Westerners. In the case of 'Safety', Westerners are negatively correlated. And in the case of 'Packages', it is relatively positively correlated with East Asians rather than Westerners.



<Figure 10> SHAP feature contribution in the Utilitarian place (East Asian (left) vs. Western(right))

국문 요지

디지털 시대가 도래하면서 온라인에서의 소비자 행동은 매우 중요한 의미를 가지게 되었다. 특히 다수의 선행연구에서 밝혀졌듯, 고관여 제품에 속하는 여행 상품은 저관여 제품에 비해 고객들의 입소문에 더 큰 영향을 받게 된다고 알려졌는데 이는 관광 산업에서 온라인 리뷰 분석의 중요성을 시사한다. 이에 발맞추어 최근 몇 년간 다수의 연구에서는 온라인 여행 리뷰와 관련된 다양한 연구들이 시행되어왔다. 하지만, 목적지 선택에 관한 연구들의 경우 여전히 리뷰 자체를 분석하는 것이 아닌, 전통적인 분석방법을 기반으로 한 리뷰어 분석이 지배적이었다. 따라서 본 연구에서는 리뷰 데이터를 이용하여 선행 연구들을 바탕으로 밝혀진 매력적인 목적지 특성들을 검증하려고 한다. 보다 세부적인 검증을 위해 리뷰어를 문화적 맥락에 따라 동서양으로 구분하였으며, 목적지는 명소의 유형에 따라 쾌락적 장소, 실용적 장소로 구분하여 분석을 진행하였다. 데이터 셋은 트립어드바이저 웹사이트에서 2021년 12월까지 작성된 모든 리뷰가 수집되었고, 다양한 텍스트 마이닝 알고리즘을 (LDA, TextRank, XGBoost, SHAP vale) 활용하여 여행객들이 목적지를 선택하는데 있어 영향을 미치는 요인과 각 요인들이 여행객의 만족도에 미치는 영향력을 분석하였다. 이후 카이제곱 검증을 사용하여 연구 결과를 검증하였으며 대응분석을 적용하여 각 특성 간의 관계를 밝히었다. 본 연구는 관광 산업에 존재하는 온라인 여행 리뷰 데이터를 다양한 알고리즘과 통계적 기법으로 풀어 제시하는 동시에 선행 연구의 내용을 검증하고 관광 산업 실무진들에게 통찰력을 제공하고 있다는 점에서 시사점이 존재한다. 관광 관련 실무진들이 그들의 목적지에 대해 관광 홍보 계획을 준비할 때 더 효율적으로 계획을 설계할 수 있도록 도움을 주고자 한다.

Keywords: 온라인 여행 리뷰, 만족도 분석, 카이제곱, 대응분석, 텍스트 마이닝, LDA, TextRank, XGBoost, SHAP value