Original Article

Applying the topic model to hotel reviews of internet sites and analyzing their results

Yoshiki Nakamura (Department of Business Administration, Aoyama Gakuin University, nakamura@busi.aoyama.ac.jp)
Nozomi Oomiya (Faculty of Advanced Engineering, Nippon Institute of Technology, n.oomiya@nit.ac.jp)

Abstract

Many people reserve and purchase travel products via internet. Travelers refer to them to gain more accurate information about specific places, the amount of time they will spend there, and so on. This paper focuses on such worldwide travel site, "A," and selects 16 hotels, which span two approaches: one from luxurious to modest, and the other, on the basis of geography, from Central Tokyo to local areas. These hotels are classified into four quadrants, and data from site A is collected on the basis of 19,224 reviews. Each data point is composed of each contributor's review text, assessment scores, and the date on which the review was published. Data points are used to analyze three things: (1) Words extracted from reviews, which are grouped according to a topic model to extract characteristics about a hotel; (2) Hotel rankings, topics, and appearance words per region; and (3) the relationships between topics and evaluations, which are quantified to consider future guidelines and services. Through these analyses, we discuss the relationship between the evaluation of a hotel and its reviews on the basis of regions and hotel rankings.

Keywords

text-mining, electronic word-of-mouth, review analysis, topic model, multiple regression analysis

1. Introduction

According to the JTB Tourism Research & Consulting Co. [2019], the percentage of people reserving and purchasing travel products via a smartphone has increased year over year. In 2018, for example, 47.3 % of all reservations were made via smartphone, as opposed to just 19.4 % in 2013. Further, 27.7 % of purchases were made for accommodation facilities, 15.8 % for restaurants, and 15.7 % for domestic tours. Similarly, Schegg et al. [2013] showed that telephone reservations are decreasing while internet-based reservations are increasing. In this context, this study focuses on travel websites. Travel websites are no longer restricted solely to travel itineraries and accommodation reservations [Nakamura and Oomiya, 2020]. Travelers now use such sites to gain more accurate information about places in order to determine the amount of time they want to spend there [Inversini and Buhalis, 2009]. Travel companies are also utilized to create marketing plans [Schmidt et al., 2008]. That is to say, review sections are especially important not only for gaining information about travel destinations and accommodations [Filieri and McLeay, 2013], but also for drawing up management strategies for the company to whom the website belongs [Chan and Law, 2006].

Ample research exists, demonstrating that travel websites can be used to gain a diversity of information and establish policies [He et al., 2017]. Among such research is that which analyzes the relationship between reviews, evaluations and satisfaction levels. Focusing on travel site reviews, Filieri and McLeay [2014] test the relationships between product rankings, information accuracy, value-added information, information relevance, and timeliness. Susilowati and Sugandini [2018] analyze a structural model from 300 questionnaires describing the causal relationship between electronic word-of-mouth, traditional word-of-mouth, perceived value, and perceived quality on the image that vacation tourists have of a destination.

At that moment, there is research that attempts to use AI techniques such as Deep Learning to predict and categorize reviews. Valdiiva et al. [2017] developed an analysis of descriptions submitted by TripAdvisor users to match between user sentiments and automatic sentiment-detection algorithms. They also discussed some of the challenges regarding sentiment analysis and TripAdvisor. Zhang and Morimoto [2017] proposes a method for recommending hotels by analyzing the text in comments using Latent Dirichlet Allocation (LDA) and extracting representative topics about hotels from the text automatically. Based on the information extracted from each topic, they make recommendations for each hotel. Xu [2018] collects data from 600 online reviews of travelers taken from Booking.com in various travel group compositions and uses Latent Semantic Analysis to identify the positive and negative factors from online reviews of travelers in various travel group compositions. Their findings indicate that not all the positive and negative textual factors mined from travelers' online reviews significantly influence their overall satisfaction. These studies, however, only take a macro perspective. In other words, they only examine reviews that have been grouped together in large data sets. However, the need for a system that is able to extract policies and issues for each hotel still exists.

This research, consequently, focuses on 16 hotels in Japan and applies a topic model to each review section. We examine the relationship between those topics and hotel evaluations in terms of regions and rankings. That is, we approach the data from a micro perspective, a method which we believe to be superior to former studies.

2. The research question and its process

2.1 The research object and its method

This research focuses on one of the worldwide travel sites, "A." A serves not only a booking the hotel, but also a tour plan,

a flight, a rent car and so on. After customer stayed the booked hotel, he/she can input various data: one is basic information as an anonymous name, a stayed date, a sex, an age, the plan and the price. The other is review as the impression and a complaint, and assessment items such as "overall," "cleanliness," "service and staff," "amenities" and "properly conditioned." Their items are 5 points of Likert scales. Some hotel replies for their review comments.

This study selected 16 hotels. We chose 16 of Japan's most famous hotels because they have more reviews than other hotels at similar price points on travel site A. In the interest of stratified analysis, we analyzed hotels from four points of view: those that were luxurious, those that were inexpensive, those located in Central Tokyo, and those located in more localized areas.

The border of the price line changes depending on the area because different areas have different commodity prices. Figure 1 presents the hotels placed in the four quadrants. Some of the facilities are in Tokyo, whereas others are located in one of five other cities including Sapporo, Sendai, Nagoya, Osaka and Hakata. Table 1 shows the hotels' number (H-ID), names, average overnight rate and user reviews. Luxurious and inexpensive hotel prices vary according to the location of the facility. Data from site A was collected 19,224 reviews from November 2014 to September 2018. Each data point is composed of each contributor's review sentences, the assessment scores and the date.

This study will try to analyze the following based on these three quadrants:

- Words extracted from reviews are grouped according to a topic model to extract characteristics about a hotel.

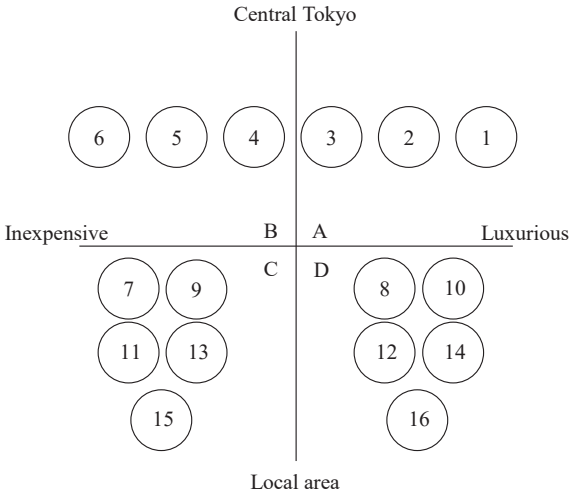

Figure 1: The standard of hotel selection
Note: Number = Hotel ID.

- Hotel rankings as well as topics and appearance words per region are analyzed.
- The relationships between topics and evaluations are quantified to consider future guidelines and services.

2.2 Data processing, morphological analysis and topic model

Data processing is explained in Figure 2. From 19,224 reviews, the 2,385 reviews submitted in Japanese are assessed. With regard to the extraction of words from the review sections, morphological analysis was conducted using Mecab [Kudo, 2010]. At that time, only nouns and adjectives were extracted. The LDA topic model was then used to obtain the frequency of words [Blei et al., 2003]. After calculating the topic model, popular topics were extracted from vast sets of

Table 1: Hotel list

H-ID	Hotel	Avg pirce ($)	Views
1	The Peninsula Tokyo	685	566
2	Imperial	330	726
3	Hotel Okura Tokyo	227	1,777
4	Park Hotel Tokyo	165	2,033
5	Shinjuku Prince Hotel	140	3,034
6	APA Hotel Shinjuku Gyoen-mae	97	1,709
7	Nagoya Ekimae Montblanc Hotel	75	1,033
8	Nagoya Marriott Associa Hotel	236	714
9	Sarasa Hotel Namba	119	793
10	Swissotel Nankai Osaka	256	3,494
11	Sendai Washington Hotel	90	394
12	Hotel Metropolitan Sendai	165	585
13	Sutton Hotel Hakata City	82	753
14	Grand Hyatt Fukuoka	219	687
15	Hotel MyStays Sapporo Station	115	1,096
16	Hotel Okura Sapporo	240	430

ID	Review date	Hotel ID	Head	Text	Value
1	2014/6/11	1	Wonderful hotel	Hotel were extremely helpful and charming	4
2	2014/6/12	2	Top of the Top	This is a true 5 star hotel. The rooms are extremely.	3
3	2014/6/12	3	Super service	My adult daughter and I had the pleasure and privilege of staying.	4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2,385		16	Super	The amenities such as the pool area is. However	5

Morphological analysis

ID	Text
1	Hotel, extremely, helpful, charming
2	Start, hotel, room, extremely.
3	Adult, daughter, pleasure, privilege, staying
.	.
.	.
.	.
2,385	Amenity, pool, area

Topic model

Topic 1 (0.5912)		Topic 2 (0.4321)		Topic 3 (0.3122)	
hotel	0.9712	room	0.9425	pleasure	0.7727
good	0.8806	staff	0.3808	hotel	0.6339
comfortable	0.7439	hotel	0.2526	room	0.1734
stay	0.6784	station	0.1614	clean	0.0245

Figure 2: Flow of data processing

sentences. This enabled us to understand the kinds of topics that belong to each sentence. In addition, the probability of the topic and the words extracted from it could be calculated (number in parentheses in Figure 2). Using these values, extracted words were clustered and the review trends per hotel along with their effect on evaluations were analyzed. Thus, in addition to the relationship between reviews and evaluations, factors such as hotel rankings and locality were analyzed from a micro perspective.

The topic model using LDA estimates the number of words that will appear in sentences from topic count (K) with initial values of $\alpha = 0.05$ and $\beta = 0.1$. Topic count (K) is estimated from the calculated values of "Coherence" and "Perplexity." According to this research, Coherence is at its highest at –2.15 when the topic count is 3, and perplexity is 190.67. Accordingly, the topic count is set at 3 (Figure 3). The probability of topics and words are produced using the above procedures.

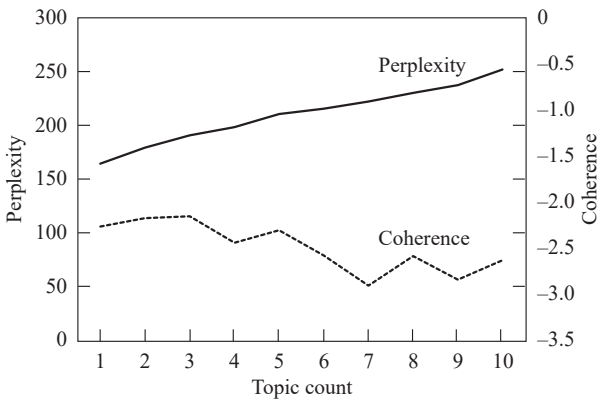

Figure 3: Coherence and perplexity shifts

3. Results of data evaluation

3.1 Basic observation

The information pertaining to each hotel is the first to be analyzed. Table 2 shows the average value of "overall," "cleanliness," "service and staff," "amenities" "properly conditioned," and price per "Luxurious-inexpensive" hotel rankings and "Central Tokyo-Local area" regions. The value for "luxurious" is higher than that of "inexpensive" for all items. In addition, the value for Central Tokyo is higher than the "Local area" value for all items. When the variance analysis was calculated, a significant difference was observed among the rows. This shows that luxurious hotels and hotels in Tokyo are more expensive and yield a higher level of satisfaction among guests.

Table 3 displays the average of evaluation value per quadrant. According to the comparison between Tokyo (Quadrant A) and Local (Quadrant D) in the context of Luxury, Tokyo had a higher value. On the other hand, according to the comparison between Tokyo (Quadrant B) and Local (Quadrant C) in the inexpensive factor, Local had a higher value for "Amenities." In other words, a hotel being located in Tokyo does not necessarily translate to consistently high evaluations across the board. When the variance analysis was calculated, a significant difference was observed among the rows.

Based on the above analysis results, we know that in Tokyo, accommodation charges and hotel evaluations are high, though this may not be the case when the factors of "Luxurious" and "Inexpensive" are included.

3.2 Analysis of the reviews by topic model

Our study proceeds to analyze the results of the topic model, which was applied to each hotel. Table 4 shows the results for

Table 2: Evaluation results and average price of accommodation per luxurious, inexpensive, central Tokyo, and local area

Level	Overall	Cleanliness	Service and staff	Amenities	Properly conditioned	Avg pirce ($)
Luxurious	4.5000	4.5750	4.5750	4.4875	4.4250	294.8
Inexpensive	4.1375	4.2000	4.2000	3.9875	4.0875	110.4
Central Tokyo	4.4000	4.4833	4.5000	4.3000	4.3500	274.0
Local area	4.2700	4.3300	4.3200	4.2000	4.2000	159.7

Table 3: Evaluation results and average price of accommodation per quadrant

Level	Overall	Cleanliness	Service and staff	Amenities	Properly conditioned	Avg pirce ($)
Lux - Tok	4.6333	4.7667	4.7333	4.6667	4.6000	414.0
Lux - Loc	4.4200	4.4600	4.4800	4.3800	4.3200	223.2
Inex - Tok	4.1667	4.2000	4.2667	3.9333	4.1000	134.0
Inex - Loc	4.1200	4.2000	4.1600	4.0200	4.0800	96.2

Table 4: Words per topic for the Peninsula Tokyo and their probability

	Topic1	0.1259	Topic2	0.1300	Topic3	0.7441
1	room	0.0219	hotel	0.0213	hotel	0.0204
2	hotel	0.0192	room	0.0171	room	0.0183
3	comfortable	0.0149	staff	0.0122	service	0.0133
4	use	0.0128	use	0.0121	people/staff	0.0098
5	staff	0.0094	stay	0.0108	men/women	0.0091
6	correspondence	0.0093	breakfast	0.0085	space	0.0088
7	good	0.0087	men/women	0.0077	staff	0.0083
8	best	0.0075	correspondence	0.0068	use	0.0081
9	men/women	0.0072	guide	0.0066	great	0.0072
10	Tokyo	0.0063	people/staff	0.0063	correspondence	0.0070

The Peninsula Tokyo. The first line is the topic number and its probability. From the second line onwards, there are ten words per topic along with the probabilities associated with them. Words shown in bold occur in several topics. They are “room,” “hotel,” “use,” “staff,” “correspondence,” “men/women,” and “people/staff.” The words in Topic 1, which are “room,” “hotel,” “comfortable,” “good,” and “best,” only occurred in Topic 1. The same applies to “stay,” “breakfast,” and “guide” for Topic 2, and “service” and “great” for Topic 3. A name was given to each topic. Topic 1 is “impression” because it contains impressions such as “good.” Topic 2 is “breakfast” because it contains “stay” and “breakfast.” Topics 3 is “service” because it contains “service,” “men/women,” and “staff.”

Similarly, topic naming was conducted for all hotels. Due to space limitation, we will now only look at the hotels that are typical and distinctive in each quadrant. For Quadrant B, we will look at Park Hotel Tokyo, which is recognized as having the best price-effectiveness in the Tokyo area [Tripadvisor, 2018]. In the case of Topic 1, the “hotel,” “use,” “station,” and “correspondence” items of convenience are ranked highest. For Topic 2, there were many place-related items such as “Shinjuku” and “place.” Because the characteristic of this hotel is that it is located in the desirable area of Shinjuku, such words are ranked highly in this topic. The topic naming was conducted as follows. Topic 1 was named “convenience” because it contains words relating to place, customer service, and accommodation, such as “use,” “correspondence,” and “business.” Topic 2 was named “hotel characteristics” because it contained items like “small” and “place.” Finally, Topic 3 was denoted “Location” because it had words like “Shinjuku,” and “Kabuki-cho.”

Quadrant C assesses the Grand Hyatt Fukuoka. This hotel is one of the most luxurious hotels in the Kyushu area. For Topic 1, “station” was the highest-ranking word, and was characterized by impression-related keywords such as “cleanliness” and “feeling.” Topic 2 contains words like “close” and “convenient,” giving the impression that the hotel is in a convenient place. Topic 3 has many “staff”-related words such as “men/women” and “staff.”

Quadrant D looks at Sutton Hotel Hakata City. This hotel is located in the same area as the Grand Hyatt Fukuoka making it possible to compare the two. Topic 1 contains many impression-related words such as “satisfaction” and “convenient.” Topic 2 lists the hotel convenience-related item of “time.” For Topic 3, there were many place-related items such as “Fukuoka,” “Canal City,” and “time.”

Table 5 similarly looks at and names topics for all hotels,

Table 5: Topic name of all hotels

Hotel ID	Topic 1	Topic 2	Topic 3
1	impression	breakfast	service
2	location	hotel characteristics	service
3	hotel characteristics	staff	service
4	hotel characteristics	cleanliness	location
5	convenience	hotel characteristics	location
6	cleanliness	impression	amenities
7	location	staff	correspondence
8	impression	cleanliness	service
9	location	cleanliness	staff
10	impression	staff	room
11	service	cleanliness	impression
12	breakfast	cleanliness	impression
13	impression	staff	room
14	impression	room	location
15	cleanliness	convenience	staff
16	location	service	staff

which will be analyzed in each of the four axes. For Quadrant A, Topic 1 shows the characteristics of each hotel while Topic 3 is consistent with service-related items. For Quadrant B, Topic 1 has a collection of hotel characteristics such as "convenience" and "cleanliness." For Quadrant C, regions vary, so a consistent assessment is difficult. However, Topic 1 contains the characteristics of each hotel, while Topic 2 has a collection of services and impressions relating to staff. For Quadrant D, Topic 1 relates to impressions, Topic 2 to rooms, and Topic 3 to services. As such, it was possible to assess many items in each quadrant rather than the mere characteristics of each hotel.

In summary, Topic 1 can be broadly classified as pertaining to items that rank highest among a hotel's priorities, while Topic 2 deals with items specific to a given hotel, and Topic 3 covers items relating to customer care, such as staff and service.

3.3 Static analysis and study of the outputs

Finally, we turn to the overall relationship between hotel evaluations and each topic number.

The correlation coefficients are shown in Table 6. The tests of non-correlation were all significant (* is 5 % significant). This tells us that there is a high correlation between evaluations and each topic.

We conducted a multiple regression analysis to identify how much factors from each topic number affect to overall evaluations (Table 7). The overall evaluation was set as an objective variable and the three topics were set as explanatory variables, with a target period of November 2014 to September 2018. The sample size for our research was 2,385. Table 7 shows outputs of the partial regression coefficient, standard partial regression coefficient, *t*-value, and *p*-value per topic together with the coefficient of determination (R2) and the Durbin-Watson ratio. R2 has high interpretability at 0.5088, making it an appropriate candidate for examination.

Table 6: Correlation coefficient results

	Value
Topic1	0.5410 *
Topic2	0.5875 *
Topic3	0.5192 *

The *t*-values of Topic 2 and Topic 3 are high, so we know that they have a large effect on overall evaluations. The case of Topic 2, in particular, is an element that is unique to hotels. Therefore, in order to improve hotel evaluations, it is vital to fully identify typically posted reviews per hotel and apply that information to hotel management. In addition, Topic 3 has a negative value. Therefore, it is crucial to reduce reviews relat-

Table 7: Evaluation and topic multiple regression analysis result

Variable	Topic1	Topic2	Topic3	Constant term	R2	DW
Partial regression coefficient	86.2727	290.1047	–399.4166	3.3857	0.5088	1.4208
Standardized partial regression coefficient	0.7153	2.6615	–2.7773			
t-value	(1.1266)	(2.295)	(–1.9991)	(4.5818)		
p-value	0.2819	0.0406	0.0688	0.0000		

ing to Topic 3.

In the case of Topic 3, the keyword "service" is recognized as a hotel's alert word for hotel 1. A given hotel can then watch the keyword according to reviews received and dependency. For example, there are reviews such as: "almost satisfied, but the service person who provided our room service was the worst. Their attitude was quite ruthless;" "service level was worse than it was during our first stay. The communication of the porter made us uncomfortable, and every staff member had a bad attitude;" and "the service level of the employees has diminished." Through these comments, suggestions regarding how to resolve issues can be proposed, including having the staff pay attention the way they talk, their attitude, and their attention to detail. This research process and output can produce a much more efficient means of identifying and resolving problems.

Based on the above analytical results, the potential of this research will now be discussed. First, applying the topic model to each hotel not only reveals review trends, but also shows the strengths and weaknesses of each region and cluster. Next, quantitatively identifying the relationship between evaluations and topics enables proposals to be made such as "being aware of service and management that attracts review words," and "applying guidelines that prevent negative words being written."

4. Conclusion

This research has conducted a review analysis of travel sites and has identified relevant information of interest to hotel management. In order to obtain robust results, we implemented an LDA topic model to the review sections and calculated the appearance probability of relevant words. By using those values, we then analyzed the linkage between specific words and hotel assessments, a hotel's location (within Tokyo), and the locality of a hotel (outside of Tokyo). For numerical consideration, the correlation coefficient was found and a multiple regression analysis was conducted in the interest of discussing future guidelines and services. We were also able to use our findings to propose ways of adjusting hotel management.

However, the following issues must also be raised: (1) Reviews should be extended to other languages besides Japanese such as English and Chinese; (2) The model should be applied to other hotel websites; (3) AI techniques other than the topic model should also be implemented; and (4) The implementation of a multiple regression analysis for each hotel results in the occurrence of multicollinearity in the case of some output results. As such, we note that the model has contradictions, which merit further consideration.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, 993-1022.

Chan, N. L. and Guillet, B. D. (2011). Investigation of social media marketing: How does the hotel industry in Hong Kong perform in marketing on social media websites? *Journal of Travel and Tourism Marketing*, Vol. 28, No. 4, 345-368.

Filieri, R. and McLeay, F. (2013). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information from online reviews. *Journal of Travel Research*, Vol. 53, No. 1, 44-57.

He, W., Tian, W., Tao, R., Zhang, W., Yan, G., and Akula, V. (2017). Application of social media analytics: A case of analyzing online hotel reviews. *Information Review*, Vol. 41, No. 7, 921-935.

Inversini, A. and Masiero, L. (2014). Selling rooms online: The use of social media and online travel agents. *International Journal of Contemporary Hospitality Management*, Vo. 26, No. 2, 272-292.

JTB Tourism Research & Consulting (2019). The state of overseas tourist travel 2018. (Retrieved January 16, 2021 from https://www.tourism.jp/en/tourism-database/studies/2018/10/state-overseas-tourist-travel-2018).

Kudo, T. (2010). MeCab: Yet another part-of-speech and morphological analyzer (Retrieved January 16, 2021 from http://mecab.sourceforge.net/).

Nakamura, Y. and Oomiya, N. (2020). An analytical examination of accommodation sales and the importance of electronic 'word-of-mouth' appraisals via internet travel sites. *Journal of Global Tourism Research*, Vol. 5, No. 1, 43-50.

Schegg, R., Stangl, B., Fux, M., and Inversini, A. (2013). Distribution channels and management in the Swiss Hotel sector. *Proceedings of Information and Communication Technologies in Tourism 2013*, 554-565.

Schmidt, S., Cantallops, A. S., and Santos, C. P. (2008). The characteristics of hotel websites and their implications for website effectiveness. *International Journal of Hospitality Management*, Vol. 27, No. 4, 504-516.

Susilowati, C. and Sugandini, D. (2018). Perceived value, eWord-of-mouth, traditional word-of-mouth, and perceived quality to destination image of vacation tourists. *Review of Integrative Business and Economics Research*, Vol. 7, 312-321.

Tripadvisor (2018). Tripadvisor travler's best choice award (Retrieved January 16, 2021 from https://www.tripadvisor.jp).

Valdivia, A., Luzón, M. V., and Herrera, F. (2017). Sentiment analysis in Tripadvisor. *IEEE Intelligent Systems*, Vol. 32, No. 4, 72-77.

Xun, X. (2018). Does traveler satisfaction differ in various travel group compositions? *International Journal of Contemporary Hospitality Management*, Vol. 30, No. 3, 1663-1685.

Zhang, Z. and Morimoto, Y. (2017). Collaborative hotel recommendation based on topic and sentiment of review comments. *Proceeding of the 9th Forum for Information and Engineering*, 6.

(Received December 8, 2020; accepted January 19, 2021)