Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

An Investigation on the User Behavior in Social Commerce Platforms: A Text Analytics Approach

Amir Arzy

Ph.D. Candidate, Department of Industrial Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: amir_arzy@atu.ac.ir

Mohammad Taghi Taghavifard*

*Corresponding Author, Associate Prof., Department of Industrial Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: dr.taghavifard@gmail.com

Zohreh Dehdashti Shahrokh

Associate Prof., Department of Industrial Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: dehdashtishahrokh@atu.ac.ir

Iman Raeesi Vanani

Assistant Prof., Department of Industrial Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: imanraeesi@atu.ac.ir

Abstract

Nowadays, the tourism industry accounts for approximately 10% of the global GDP, while it only contributes 3% of the economy in Iran. Since the pressure of US sanctions increases day after day on the Iranian economy, the necessity of paying attention to this industry as a source of foreign currency is felt more than ever. The purpose of this research is to analyze the reviews of users of social commerce websites by using a combination of text mining and data mining techniques. For this purpose, the database of TripAdvisor website (TripAdvisor.com) was evaluated, and all profile information of users who commented on hotels in Iran was collected. These comments on all the content of the website, such as hotels, restaurants, and attractions, were then extracted and analyzed. The optimal number of clusters was considered four clusters by calculating the Davies-Bouldin index, namingly water therapy tourists, boutique hotels style and Iran urban tourists, travelholics and food tourists, business and health tourists. Every single cluster possesses unique attributes and features. Afterward, the association rules were further identified for each cluster according to the characteristics of each cluster and the information in the users' profiles. Finally, a solution is proposed to increase the

participation of the users on the website, and targeted promotional plans are expressed in accordance with the well-known features of each cluster.

Keywords: Social commerce, TripAdvisor, Social media, Text mining, Data mining.

DOI: 10.22059/jitm.2020.296648.2458

Introduction

In today's marketing science, experts focus on the word of mouth marketing and encourage organizations and companies to perform this method more. With the development of Web 2.0 and the rise of social commerce websites, word of mouth marketing has been converted to an electronic form, by which the users on social networks express their point of views and experiences on products, services, and companies; thus, their reviews may affect other people on those social networks to buy/not buy the products or services.

The global number of internet users exceeded 4.4 billion in the year 2019, and the growth of which has been shown compared to the previous year according to E-marketer forecast. Currently, the internet penetration rate is approximately equal to 57%. It is estimated that more than 5 billion people worldwide will have access to the Internet by the end of 2020. Inexpensive cell phones and their connection to the Internet are expanding in developing countries with no sufficient infrastructure and facilities to access the Internet. Monica Pierre, the senior analyst at E-Marketer, said: "When high-growth markets reach saturation levels in terms of Internet access, it is developing markets that can offset this growth rate" (E-Marketer, 2019).

According to the survey of Korean institute in 2018, 45% of consumers were influenced by other users' comments, and 20.6% of consumers expressed their opinions or recommendations in online communities, blogs, and online shopping markets and shared their opinion with others (National Internet Development Agency of Korea, 2018). Also, some researches demonstrate that potential customers are more probable to read other people's comments and opinions through the online rating of websites, and they attend to online ratings more than using the information provided by the companies (Ridings & Gefen, 2014).

The present article analyzes user reviews on social commerce websites on the basis of text mining and data mining techniques on TripAdvisor. In the following, the authors describe the problem and then express the topic, the significance of the study, and the research background. Afterward, the research methodology and the data analysis were described, and in the final section, the results were discussed.

Nowadays, customers not only enter the online shopping sites for shopping but also bring in all their social networks for this purpose (Lu et al., 2010). This collaborative environment has transformed users from passive behaviors to active content producers on the web (Zwass, 2010; Hajli, 2012). In this interactive environment, customers not only purchase products or services but also create content that leads to a two-way relationship for the seller and consumer (Park et al., 2007).

In online shopping, customers cannot sense the products such as by touching or smelling closely; therefore, other users' opinions become increasingly important, particularly when others are expressing their opinions based on personal experience with the consideration of a product or services. These comments and ratings are essential parts of potential customers' behavior (park et al., 2007).

Social commerce is a paradigm, moved business practices, and opened new windows to information systems research. According to a study conducted by Curty and Zheng, social commerce refers to the business activities that are facilitated by social media services. Many brands appear on these social websites to advertise, promote, and perform word-of-mouth marketing (Curty & Zheng, 2011).

Considering the abovementioned issue and explaining the importance of user reviews on social commerce websites, it is evident that the main purpose of this research is to investigate the behavior of users in social commerce websites in the field of the tourism industry.

The main problem is that many users only play a passive role in the social commerce websites and do not participate in them by writing reviews about their experiences to help other users increase the usefulness of the website and select proper services or products. Furthermore, there are not many studies on the clustering of users according to their reviews and comparing the reviews with their participation based on their profile information. According to this combination, the present study is performed to provide some solutions to increase the participation of each cluster individually. The authors applied data mining and text mining techniques on this hidden treasure to make the ultimate benefits for social commerce websites and domain tourism activists. The appropriate action can be taken in this industry according to the findings.

Literature Review

Social Commerce

Some researches show that potential customers are more interested in other people's opinions and suggestions compare to the information and explanations provided by product makers or

service providers (Turban et al., 2017). On the other hand, this issue is also increasing in Iran, and the importance of studying this subject is more practical with the growing number of Iranian social business sites such as Digikala.com, Alibaba.ir, Snapptrip.com, etc.

A significant portion of the business community has entered the e-commerce marketplace with this massive volume of business opportunities provided by e-commerce markets, and online shoppers are applying their social capital to increase their purchase volume and get the lowest possible price (Leitner & Grechenig, 2008). Social commerce may be used for business-to-consumer, business-to-business, consumer-to-consumer, and employee-to-employee (Saundage & Lee, 2011).

For the first time, Yahoo introduced the term "Social Commerce" to describe its online participatory shopping tools and user ratings. The company desired to create a community of buyers who would rate the products and share their experiences, and the information produced by that community of users would eventually become available to other buyers.

No precise definition is available for social commerce because it has different meanings for different people. It is generally defined as a subset of e-commerce that uses social networks to support social interactions to buy and sell products and services online (Mangold & Faulds, 2009).

Text and Data Mining Research on Comments

A review of some related research in text mining and data mining in tourism is discussed in the following.

Chang, Hsu, Cheng and Chung (2015) identified the fictitious reviews from real reviews among users about hotels. They claim that as consumers grow in importance to other customers and influence their decision-making, they are more probable to decide to purchase the fake and self-made ideas from business owners to increase their sales or cover their problems. This behavior burns customers and businesses and has a damaging effect on customers' trust. There are many influential factors when it comes to the decision to buy and the probability of using products and services, that social media is one of them. The important point is that the origin of all these influences is based on trust (Chang et al., 2015).

In order to perform this action, text-based speech modeling was used according to the extraction of three important descriptive word indices, the numerator, and the ratio of inactive words. Also, a rumor detection model was used to distinguish between a real and a fake comment. The study also uses comments on the TripAdvisor website.

In another study, Dickinger and Lalicic (2016) assessed important tourism destinations using the reviews of users on the TripAdvisor website. The present research focused on the

feelings expressed in users' perceptions of tourism destinations, such as restaurants, hotels, "things to do," and the achieved significant results. The research claims that the dimensions of sophistication, excitement, ruggedness, sincerity, and competence are far better reflected in social media than conventional research. Feelings of anger, hatred, and discomfort are also more evident on social media. They used the K-mean text clustering method by Wordstat software to analyze customer reviews (Dickinger & Lalicic, 2016).

Qi, Li, Zhu and Shi (2017) Surveyed 16,000,000 user reviews on Ctrip.com (China's largest hotel information site) at 70 five-star hotels by text mining techniques. The results indicated that most of the hotel's interior complaints focused on the three factors of air conditioning, noise, and humidity, which have been achieved through text clustering. Also, it was discovered that the ranking of the quality of indoor hotels indirectly impacts the hotel business. The most important thing about customer retention is to create a sense of satisfaction. This issue has led companies to conduct further studies on identifying customer behavior and the factors affecting their satisfaction in the internet environment (Qi et al., 2017).

Afrizal, Rakhmawati and Tjahyanto (2019) surveyed user reviews of tourism products using opinion mining. By using filtering techniques, they could improve feature extraction and the results of opinion classification. The focus of their research was on improving the automated methods of filtering textual data. They used word weighting techniques such as TF-IDF, BM25, etc. to improve the results of feature extraction and, consequently the opinion mining and finally TFID as the best method for weighting text data (Afrizal et al., 2019).

Annisa and Surjandari (2019) conducted a study on one of the free tourism zones of Indonesia. The hotels in this area are considered as one of the tourist experiences of the area and as a reference point for people's comments on social media. They claimed that the results of this research could give feedback to hotel owners on whether their business is performing well and how they can improve their businesses. By using Latent Dirichlet Allocation, they surveyed the content of user reviews of hotels in the area and finally discovered eight important hotel topics, for which there were typical conversations between hotel owners and tourists (Annisa & Surjandari, 2019).

Villeneuve and O'Brien (2020) surveyed users' opinions about Airbnb resorts. They surveyed 1.35 million reviews of Canadian users who had experiences of living in resorts. They focused on the major domestic grievances of these resorts and their seasonal understanding of such complaints. Their results revealed that about 5% of comments were about complaints about the indoor environment of these resorts. Their method was a combination of quantitative and qualitative methods, while conventional text mining methods had limitations on accurate word interpretation and precise relation of words. They used text

clustering by the K-mean method to categorize customers' complaints on indoor quality problems, and cluster reviews to 4 main complaints. Afterward, they described particular solutions to resolve the problems (Villeneuve and O'Brien, 2020).

Materials and Methods

As discussed in the literature review, several studies have been conducted on text clustering of customer reviews in the tourism industry, but none of them have employed text clustering for studying customer behavior on participating in social commerce websites. The present research simultaneously studied text clustering of user reviews and data derived from users profiles. The present study applied association rules for the first time to describe each user's cluster attributes and recommend some marketing solution to increase users' participation and selling volume for tourism-related businesses.

In order to collect the information of this study, the method of library-based content analysis, web-based content analysis, and the reviews and opinions of users of the TripAdvisor website were used. Firstly, all users who commented on one of the hotels in Iran were identified. Afterward, their profile information and all their comments on items such as hotels, restaurants, and places of interest were extracted. All of these data were collected by a web based program and Python codes were used to extract and save profiles data in an excel file and each user's comments in a separate text file. After extracting information and comments, data and texts were analyzed by employing data mining and text mining tools. This sampling was made because Iran is one of the most important destinations for tourism due to its rich culture and antiquity, and it can assist in planning more tourism and attracting foreign exchange earnings for the country at this critical juncture. The total number of 11,040 users and over 470,000 comments were extracted. Data analysis and text mining techniques and clustering methods were used to analyze the data in the study database. In the present study, RapidMiner and Clementine software were used to analyze the data in the database.

CRISP-DM is one of the methods used in this research. As can be seen in Figure 1, this method possesses six steps (Wirth & Hipp, 2000):

- **Business Understanding**: At this stage, all the requirements and business conditions must be thoroughly evaluated. The ultimate purposes of the business are determined through conversation with senior business executives. Business understanding involves setting business goals, evaluating positioning, setting data mining goals, and preparing a project plan.

- **Data Understanding:** This section begins with the initial data collection step and will eventually lead to an introduction to the data. Data comprehension includes the steps of

collecting basic information, data description, data discovery, and determination of data quality.

- **Data Preparation:** The data preparation step covers all the activities required to create the final dataset. In this final dataset, the data is fed into the modeling tools. This step includes data selection, data cleaning, data construction, data integration, and data formatting.
- **Modeling:** At this step, different modeling techniques are selected and applied. Furthermore, their parameters are calibrated to achieve optimal results. This stage includes modeling techniques, test design production, model development, and model evaluation.
- **Evaluation:** At this stage, models that determine the high quality of data analysis should be developed. This step includes evaluating the results, reviewing the process, and identifying the next steps, at the end of which the necessary decisions must be made based on the data mining results.
- **Deployment:** At this step, the outputs created to improve the business are used. This step includes deployment planning, control and maintenance of planning, preparation of the final report, and project review.

Figure 1 shows the CRISP-DM process used in the study.

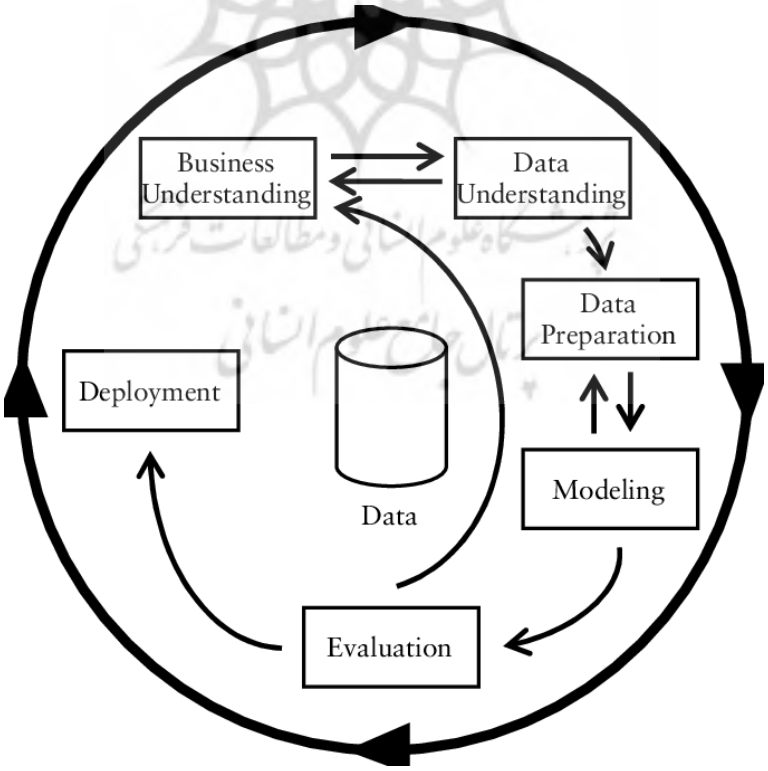

Figure 1: CRISP-DM Life Cycle (Wirth & Hipp, 2000)

The CRISP-DM cycle used in this studyis indicated in Figure 2:

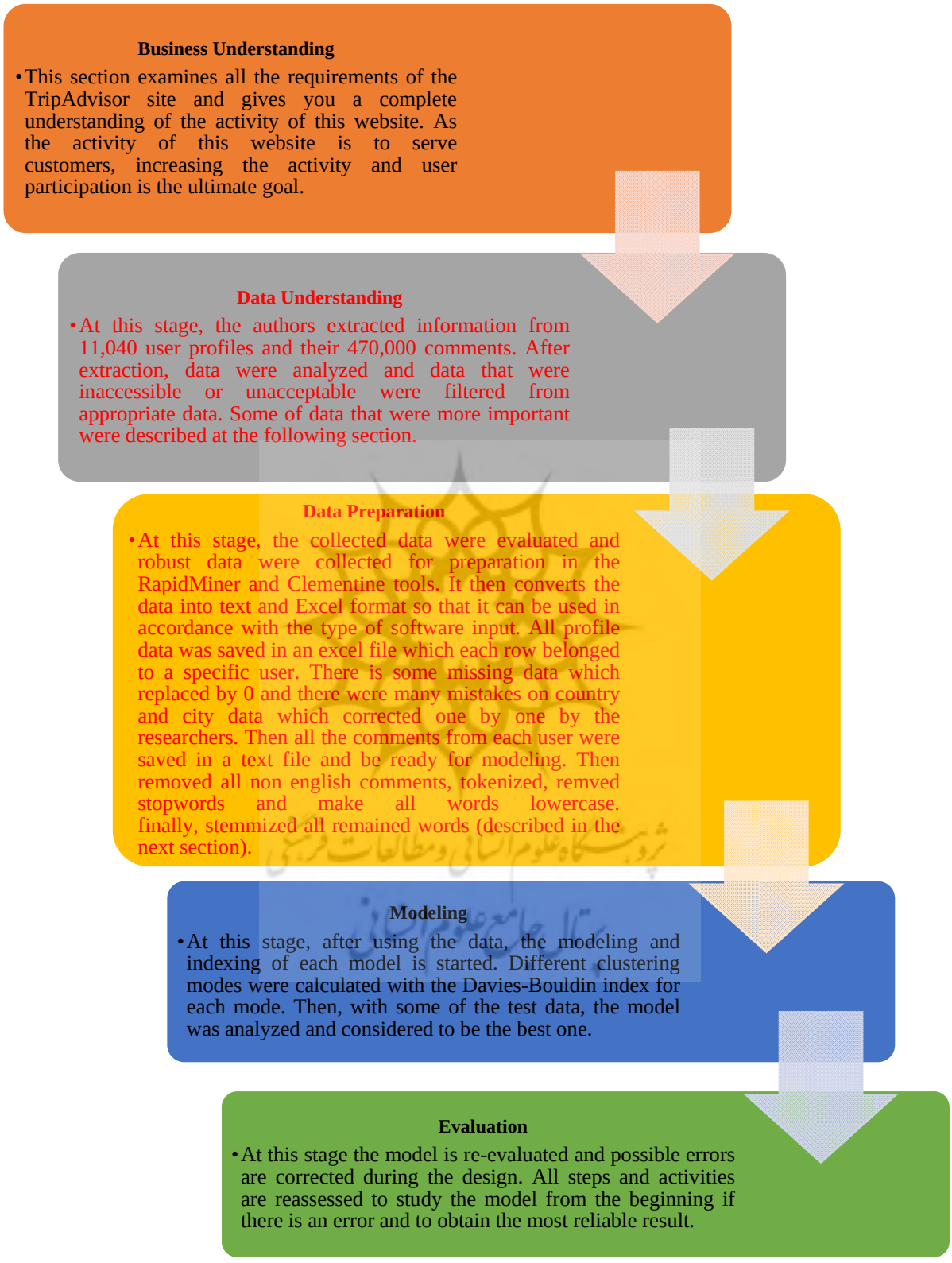

Figure 2. CRISP-DM cycle used in this study

Data Pre-Processing

In the data preparation stage, all the data were collected from users' profiles and were later reviewed; 0 was used for missing values in each variable. Also, there were many mistakes in information about city and country because of the wrong structure of the Tripadvisor website, which were individually corrected by the authors.

For preparing user reviews for the text mining process, the authors performed several steps after saving each user review in a separate text file as follows:

- **Lower casing:** The method of lower casing analysis is used to convert all input texts to lower case words so that, words "Yes," "yes," and "YES" are treated in the same way.
- **Removal of punctuations:** Removing punctuations from the text data (characters like: !@&*%$); this is a text standardization to see "Yes" and "Yes!" in the same way.
- **Removal of stop words:** Stop words are regularly occurring in each language, such as "the," "a," "in," etc. They can be illuminated from the text because they do not provide valuable information for more analysis.
- **Removal of frequent words:** As stop words were removed, there are frequent words (such as "yes," "no," "please") that are only used frequently without any significant sentiment intentions. In this stage, these kinds of words were removed.
- **Correction of spell:** Absolutely, there are many grammatical and spelling mistakes in user reviews, which should be corrected.
- **Stemming of words:** This is the process of reducing words to their root forms. For instance, if there were two words "friend" and "friendly," the stemming will stem the suffix to see them as the same word as "friend".

Defining the Variables Used in the Research

- Reviews: The number of user reviews on various tourism items on the website.
- Photos: The number of photos of tourist sites uploaded by the user on the website.
- Forums: The number of forums in which the user has participated and discussed.
- Contribution: The total number of videos, comments, photos, and reposts by each user.
- Level: The level on the TripAdvisor website is calculated based on the points that each user earns (300 points level 1, 500 points level 2, 1000 points level 3, 2500 points level 4, 5000 points level 5, and 10,000 points level 6).

- Helpful Votes: The number of votes that users give to the other user's comment.
- Readership: The number of times other users have read a user's comment.
- Passport: A passport is added to each city in terms of hotels, restaurants, places of interest, etc.; this is the index of the number of cities that a user has commented on their tourist sites.
- Followers: The number of people who follow the user on the TripAdvisor website.
- Country: It is referred to as the user's country of residence.
- Total miles traveled: The total distance the users have traveled from their location to the tourist sites, where they have commented on.

Discussion

Clustering the Users Based on the Content of Reviews

The Davies-Bouldin index was used to determine the optimal number of clusters for K-Means clustering. In this way, users' reviews were divided into 2, 3, 4, 5, 6, 7, and 8 clusters by RapidMiner software. Then the Davies-Bouldin index was calculated for each of these clusters, and finally, the optimal cluster number was selected based on the results of four clusters. The value of this index was equal to 0.002 for the 4 clusters, which was the lowest value compared to the other clusters. Figure 3 shows the portions of each cluster.

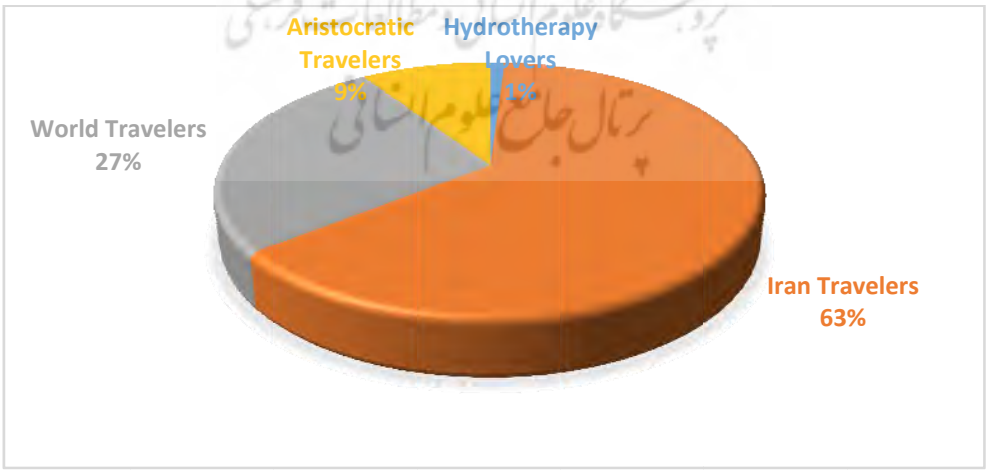

Figure3. Clusters' information

Table 1 presents the results of the clustering of user reviews based on the K-Means algorithm and four clusters. This table lists 35 more repetitive words for each cluster.

Table 1. The results of clustering user comments based on the K-Means algorithm

Term	Cluster 1 Centroid	Term	Cluster 2 Centroid	Term	Cluster 3 Centroid	Term	Cluster 4 Centroid
sarein	0.817640	shiraz	0.033881	was	0.055370	espinas	0.206039
royal	0.199465	yazd	0.032131	but	0.051606	tehran	0.157631
ardabil	0.174793	isfahan	0.030212	restaur	0.050204	gulf	0.077803
park	0.106277	tehran	0.026874	dubai	0.049469	khomeini	0.074282
hydrotherapi	0.038551	kashan	0.024341	not	0.048749	palac	0.062113
sabalan	0.031510	restaur	0.018326	you	0.048276	imam	0.059710
miner	0.025982	was	0.017550	the	0.044982	persian	0.055936
decemb	0.021777	you	0.016571	were	0.043510	markazi	0.053101
sareyn	0.020388	tabriz	0.014402	that	0.040216	novotel	0.045851
royalpark	0.019403	ameriha	0.014215	there	0.038609	airport	0.036626
famili	0.018366	good	0.013927	europ	0.037758	international	0.034119
complex	0.014653	not	0.013819	for	0.037475	thank	0.031277
sauna	0.014365	but	0.013744	are	0.036480	ibis	0.025702
water	0.014006	the	0.013277	which	0.036474	business	0.021644
Spa	0.013439	visit	0.013241	visit	0.035876	grand	0.018567
beauti	0.012581	place	0.013232	had	0.035655	septemb	0.016314
relax	0.012395	saray	0.013215	asia	0.035457	veri	0.015888
best	0.011975	nice	0.013107	have	0.033659	was	0.015763
therapi	0.011448	great	0.013054	from	0.033416	service	0.015569
stylish	0.011289	veri	0.012935	our	0.033393	juli	0.015327
khansalar	0.011262	are	0.012548	this	0.032738	profession	0.014806
massag	0.011108	europ	0.012458	food	0.032649	good	0.014374
spring	0.009951	this	0.012440	good	0.031807	staff	0.013814
mountain	0.009905	kish	0.012394	great	0.031345	nice	0.013770
treatment	0.009502	for	0.012275	they	0.031198	excel	0.013635
great	0.009404	mashhad	0.012233	unit	0.030909	great	0.013563
Salt	0.009062	boutiqu	0.012141	thing	0.030572	august	0.012936
pool	0.009030	there	0.012115	experi	0.030566	kind	0.012882

Term	Cluster 1 Centroid	Term	Cluster 2 Centroid	Term	Cluster 3 Centroid	Term	Cluster 4 Centroid
pichagh	0.008889	dubai	0.012110	place	0.029391	help	0.012077
cave	0.008706	experi	0.011985	one	0.029078	best	0.011990
suggest	0.008646	morshedi	0.011949	room	0.027766	friend	0.011780
place	0.008488	august	0.011860	well	0.027270	for	0.011278
moment	0.008306	septemb	0.011853	emir	0.026798	special	0.010547
Was	0.008292	food	0.011618	would	0.026721	especi	0.010369
khansalar	0.011262	were	0.011550	veri	0.026325	hospital	0.010093
massag	0.011108	are	0.012548	this	0.032738	profession	0.014806
jacuzzi	0.008159	europ	0.012458	food	0.032649	good	0.014374

Results of Clustering

The following results can be derived from the information presented in Table 1:

Cluster One

The total number of 110 users is associated with this cluster, which comprises about 1% of the total population. Considering the contents of these comments, it is clear that the main purpose of these users' trips was hydrotherapy and, in particular, the natural hot springs of Sarein Ardabil. These people have traveled to Ardabil and Sarein, and both have a keen interest in water recreation, such as the sauna, Jacuzzi, and pool, as well as water treatment issues in Ardebil's famous hot springs. Most of these users share their experience of Royal Park Water and Sarein Therapy and are also interested in massage and relaxation. Given the high consistency of the word "great" in their comments, these users are delighted with their tourism experience and are thrilled. These users have mostly opted for spring or January, and in the end, they are interested in the family-style trip. By these attributes we named this cluster as Water Therapy Tourists.

Cluster Two

The total number of 6981 users is associated with this cluster, which covers 63% of the total population. The reviews of this group of users are focused on tourism from different cities of Iran. They have visited Shiraz, Yazd, Isfahan, Tehran, Kashan, Tabriz, Kish, and Mashhad, respectively. The important thing about this cluster is that these people prefer to go around the restaurants; thus, the word "restaurant" is prevalent. Many of these users share a stay at the Saraye Ameriha Boutique Hotel and the Morshedi House in Kashan, both boutique and traditional Iranian hotels. Three words of "great," "good," and "nice" are visible in their

comments that indicate their high level of satisfaction. These users usually choose September, October, and November for their journey. As described above, we named this cluster as Boutique Hotels and Urban Tourists.

Cluster Three

The total number of 2936 users is associated with this cluster, which covers 27% of the total users. The main focus of this cluster is on restaurants. The important point about the tourist destinations of this cluster is users from Dubai, European countries, and Asian countries. In fact, they are professional tourists who have also visited Iran. There is some satisfaction in their comments, but given the placement of the term "but" in the second sequence of words used in their comments, some criticism can be felt. Due to the tourist attractiveness of these people and their stay in the hotel, they mostly expressed their opinion about the hotel rooms. Finally, it should be noted that the users of this cluster did not make much difference to travel in different months of the year. As explained for this cluster, we named it as Travelholics and Food Tourists.

Cluster Four

The total number of 1013 users are associated with this cluster, covering approximately 9% of the total population. The most prominent feature of this cluster is the subscription review of Espinas Palace Hotel, the Espinas Gulf Hotel, and the Novotel-Ibis Hotel. The first two hotels are one of the most luxurious and expensive hotels in Iran, reflecting the high financial level of users of this cluster. A large part of the cluster has been staying at these hotels due to business-related issues, and this is becoming increasingly apparent as the Novotel Hotel at Imam Khomeini International Airport is one of these destinations. Some users of the cluster have also commented on the hospital, which could be related to health tourism. July, August, and September have been preferred times of the year for travelers. The type of service and attitude of the staff is also important for these travelers, and they generally expressed their satisfaction and pleasure in their comments. These clusters have preferred friendly trips more than family trips. Regarding the mentioned features, we called this cluster as Business and Health Tourists.

Association Rules Derived from Users' Content Clustering

Table 2 shows the association rules extracted from the users' profiles and the results of the users' clustering based on the content of their comments.

Results of Association Rules

The following results can be deduced from the information given in Table 2:

Table 2. Association rules derived from users' profile features and clustering based on the content of comments

Row	Consequent	Antecedent	Support%	Confidence%	Row	Consequent	Antecedent	Support%	Confidence%
1	Cluster = 2.0 and Junior Photographer = 5.0	content = cluster_2	0.15	100	16	world traveled > 0.045 and reviews > 20.5	content = cluster_2	27.03	75.67
2	Cluster = 2.0 and photos > 66.5	content = cluster_2	0.14	100	17	Year > 2013.5 and Year < 2017.5 and reviews < 14.5	content = cluster_1	29.66	87.82
3	country = Belize	content = cluster_3	0.02	100	18	Cluster = 4.	content = cluster_2	6.53	80.05
4	country = Cuba	content = cluster_3	0.02	100	19	country = Syria	content = cluster_3	0.05	66.67
5	Cluster = 2.0	content = cluster_2	0.19	90.91	20	country = Canada and Month = Jul and total miles traveled < 7341.5	content = cluster_0	0.02	100
6	Helpful Votes < 8.5 and reviews > 1.5	content = cluster_1	34.95	89.71	21	Year > 2013.5 and Contributions < 25.5 and Contributions > 1.5	content = cluster_1	29.65	87.69
7	Cluster = 4.0 and Readership > 27500.0	content = cluster_2	5.14	89.33	22	forum < 0.5 and reviews < 20.5 and Contributions > 1.5	content = cluster_1	41.22	87.58
8	Cluster = 4.0 and Helpful Votes > 26.5	content = cluster_2	5.14	88.67	23	photos < 8.5 and Helpful Votes < 12.5 and Readership > 300.0	content = cluster_1	39.49	87.51
9	Cluster = 4.0 and reviews > 32.5	content = cluster_2	5.21	88.16	24	reviews < 14.5	content = cluster_1	61.5	80.67
10	reviews > 28.5	content = cluster_2	27.56	74.15	25	Luxury Hotel Expert = 0.0 and Senior Contributor = 0.0 and Expert Photographer = 0.0	content = cluster_1	62.58	80.62
11	world traveled < 0.035 and Year < 2017.5 and reviews < 10.5	content = cluster_1	35.01	88.06	26	Luxury Hotel Expert = 0.0 and Senior Contributor = 0.0 and Top Contributor = 0.0	content = cluster_1	62.99	80.56

Row	Consequent	Antecedent	Support%	Confidence%	Row	Consequent	Antecedent	Support%	Confidence%
12	Luxury Hotel Expert = 0.0 and Senior Contributor = 0.0	content = cluster_1	62.99	80.56	27	country = 0	content = cluster_0	87.93	100
13	Contributions < 20.5	content = cluster_1	59.79	80.29	28	country = Canada	content = cluster_0	1.72	100
14	Contributions > 30.5	content = cluster_2	33.4	67.95	29	country = Iran	content = cluster_0	1.05	100
15	Cluster = 1.0	content = cluster_1	93.27	66.64	30	Level = 0.0	content = cluster_0	0.35	79

Cluster One (Water Therapy Tourists)

Due to the small number of users in this cluster, the association rules were also extracted once individually for this cluster to get more insight. The major national users of this cluster were Canadians, Iranians, and unknown countries. These clusters are also low-level in terms of participation and have little involvement in both photo-sharing and review writing. Users of this category have been reluctant to stay in luxury hotels, and their focus has been only on the use of spa and mineral water.

Cluster Two (Boutique Hotels and Urban Tourists)

Although different cities in Iran have been targeted by this cluster of travelers, more than half of them fall into the category of low-travelers. These users are not staying in luxury hotels and are also not professional in photo-sharing or review writing. Over 80% of them have less than 14 comments on the website, and their turnout is less than 20. Furthermore, more than 90% of them had less than nine helpful votes. Users of this cluster also possess fewer than nine photos, 13 helpful votes, and 132 contributions. They are also relatively inactive in the forums. Most of them joined TripAdvisor between 2013, 2014, 2015, and 2016, and 88% of them traveled to less than 0.035% of destinations of the world.

Cluster Three (Travelholics and Food Tourists)

Users of this cluster are highly professional in photo sharing, with more than 67 photos, and they possess the title of junior photographer. In addition, nearly all of the target community's active people are in this cluster. These people mostly have more than 28 comments. Also, several users have stayed in luxury hotels with high turnout, over 32 comments, and 26

helpful votes. Users of this cluster had a travel rate higher than 0.045%, more than 20 reviews, and more than 30 contributions that indicate their globality and professionalism in travel and tourism.

Cluster four (Business and Health Tourists)

Most of the rules extracted from this cluster focus on the type of their country. Almost all users in Cuba, Belize, and Syria belong to this cluster. The important thing regarding the rules of this category is the relationship between the country and the month of travel, which is interesting, which is summarized in Table 3.

Table 3. The relationship between the month of travel and countries in the cluster four

Row	Month	Country	Row	Month	Country
1	Sep	Argentina	5	Dec	Iraq
2	Jun	Azerbaijan	6	Apr	Qatar
3	Sep	Cyprus	7	March	Greece
4	Oct	Finland	8	March	Hungary

Analysis of Findings

In this section, the authors propose some recommendations for each cluster of users to improve user engagement on the website. Besides, the proposed clustering provides insight for decision-makers in the different sectors of the tourism industry and allow them to target each cluster of users by a marketing plan that better fits their needs and expectations, which reduces the marketing cost and increases the overall profit.

Cluster One (Water Therapy Tourists)

- Users in this cluster manifest a deep interest in hydrotherapy (water cure). Therefore, the owner of such businesses can easily target and serve them.
- Most of the users in this cluster use the website only in the spring season. Therefore, business owners must offer some promotions based on it, so that they have a steady demand throughout the year.
- In addition to the owners of hydrotherapy facilities, the owners of restaurants, hotels, and other businesses in the cities of Sarin and Ardebil can benefit from this market by providing family-friendly promotions.
- Other relaxation and massage businesses in Iran can invest in these users as potential future customers.

- Canada is an excellent target for promoting Iran's hydrotherapy sites.
- Given the low engagement of these users, it would be beneficial if one can somehow encourage them to share their photos and experiences with others.

Cluster Two (Boutique Hotels and Urban Tourists)

- Users in this cluster usually prefer to stay in boutique hotels inspired by local culture and foods. Such hotels should offer different promotions to the users in this cluster.
- Restaurants are one of the main tourist destinations, especially for the users in this cluster. Shiraz, Yazd, Isfahan, Tehran, Kashan, Tabriz, Kish, and Mashhad can benefit from these users.
- Due to the large size of this cluster and the low user engagement on the website, it is desirable to offer a variety of promotions to whom they share their experience and photos through the website. Then, it is expected to observe a significant jump in the number of submitted reviews on the website.

Cluster Three (Travelholics and Food Tourists)

- Users of this cluster are travel enthusiasts. As a result, it is expected from them to notice more details when traveling and compare the services they received in different destinations. These users usually share their travel stories in different social media, and consequently, they can influence hundreds to thousands of other users. Any dissatisfaction for this group is expected to become viral and to impact the local businesses negatively.
- If this group of users remains satisfied with the strategy of word of mouth marketing might provide a golden opportunity for promoting and prospering local businesses.
- Luxury hotels are among the main destinations of these users. Owner of the luxury hotels should target these users by offering some incentive packages.

Cluster four (Business and Health Tourists)

- The users in this cluster prefer to stay in one of the Espinas hotels and resorts. Espinas hotels can serve a significant portion of these users by targeting them and performing the proper marketing plan.
- Some of these users are businessmen/businesswomen who have been staying at the Imam Khomeini Airport Novotel Hotel. The other hotels in the vicinity of the airport can target this group of users in their future business trips.

- According to Table 4, a proper marketing plan can be implemented for the residents of each country according to the given timetable.
- Considering the fact that this cluster also includes health tourists, hospitals and treatment groups can attract them for their services.
- Also, the website can offer them some bonuses to share their experiences about their treatment tourism and also recommend them other related hospitals for following their treatment process.

Conclusion

The present study simultaneously examined the content of users' comments and their profile information, while other surveys such as Turban, Villeneuve, and Annisa focused individually on their reviews. By clustering users and analyzing the behavior of them within each cluster, the authors provided insight for decision-makers in the tourism industry. The targeted users on the TripAdvisor website have been clustered into four groups, namingly Water Therapy Tourists, Boutique Hotels and Urban Tourists, Travelholics and Food Tourists, Business and Health Tourists. The Water Therapy Tourists cluster mostly traveled to Sarein and Ardabil, and the purpose of their travel was hydrotherapy. The Boutique Hotels and Urban Tourists cluster traveled to the popular Iran cities like Shiraz, Yazd, Isfahan, Tehran, Kashan, etc. They were located in the low travel group and preferred to stay in cheap or boutique hotels. The Travelholics and Food Tourists cluster were professional tourists who travelled around the world. They were also food tourists and people interested in different cuisines. Business and Health Tourists were the last cluster which used Espinas hotels group and also Novotel hotel. This cluster had many reviews about hospitals and health tourism in general. Then the profile attributes of each cluster were determined by applying association rules and analyzing users' comments and profile information.Afterward, some solutions were presented based on the characteristics of each cluster to maximize the participation of each category of users and increase the profit of companies which are active in the tourism industry.

Studying user profile information and the content of their reviews helped the authors understand user participation behaviors more accurately and get an insight into how user participation in websites reflects their point of view with a specific interest in tourism. Then, social commerce websites can plan an extensive range of marketing tips to encourage them to participate more and more in their websites. Finally, it should be noted that user engagement in social media is important from two points of view; first for the benefit of social media and the other for the way of increasing user engagement in the media. These two dimensions reinforce each other as a loop. In other words, the more social media users are involved, the more the results are credible, in addition to extracting more accurate results from the same

media. Evidently, the higher the media credibility, the higher the engagement of users in that medium, and so it will become the circular trend.

The present study focused on the TripAdvisor website, and it can applied to other tourism websites such as booking.com or hotels.com, which can specifically focus on hotels. Furthermore, we used the K-mean method for clustering reviews, while other researchers can use different methods and compare them regarding cluster efficiency. Finally, several studies have been condcuted on users' intention to participate in the social commerce websites, namingly Hajili et al. (Hajili et al., 2015) and Cho and Son (Cho & Son, 2019). The combination of the results of this research with factors of users' intention to participate described by the mentioned researches will reveal the new opportunities for a wide range of marketing studies.

References

Afrizal, A. D., Rakhmawati, N. A., &Tjahyanto, A. (2019). New Filtering Scheme Based on Term Weighting to Improve Object Based Opinion Mining on Tourism Product Reviews. Procedia Computer Science, 161, 805-812.

Annisa, R., &Surjandari, I. (2019). Opinion Mining on Mandalika Hotel Reviews Using Latent Dirichlet Allocation. Procedia Computer Science, 161, 739-746.

Chang, T., Hsu, P. Y., Cheng, M. S., Chung, C. Y., & Chung, Y. L. (2015, June). Detecting fake review with rumor model—Case study in hotel review. In International Conference on Intelligent Science and Big Data Engineering (pp. 181-192). Springer, Cham.

Cho, E., & Son, J. (2019). The effect of social connectedness on consumer adoption of social commerce in apparel shopping. Fashion and Textiles, 6(1), 14.

Curty, R. G., & Zhang, P. (2011). Social commerce: Looking back and forward. Proceedings

Dickinger, A., &Lalicic, L. (2016). An analysis of destination brand personality and emotions: A comparison study. Information Technology & Tourism, 15(4), 317-340.

Emarketer. (2019). Internet to Hit 3 Billion Users in 2015, http://www.emarketer.com/Article/Internet-Hit-3-Billion-Users-2015/1011602.

Hajli, M. N. (2012). An integrated model for e-commerce adoption at the customer level with the impact of social commerce. International Journal of Information Science and Management (IJISM), 77-97.

Hajli, M. N., Shanmugam, M., Powell, P., & Love, P. E. (2015). A study on the continuance participation in on-line communities with social commerce perspective. Technological Forecasting and Social Change.

Leitner, P., &Grechenig, T. (2008). Collaborative shopping networks: Sharing the wisdom of crowds in E-commerce environments. BLED 2008 Proceedings, 21.

Lu, Y., Zhao, L., & Wang, B. (2010). From virtual community members to C2C e-commerce buyers: Trust in virtual communities and its effect on consumers' purchase intention. Electronic Commerce Research and Applications, 9(4), 346-360.

Mangold, W. G., &Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. Business horizons, 52(4), 357-365.

National Internet Development Agency of Korea. (2008). Social Software: Beyond Consumer, Go Enterprise.

Park, D. H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. International Journal of Electronic Commerce, 11(4), 125-148.

Qi, M., Li, X., Zhu, E., & Shi, Y. (2017). Evaluation of perceived indoor environmental quality of five-star hotels in China: An application of online review analysis. Building and Environment, 111, 1-9.

Ridings, C. M., &Gefen, D. (2004). Virtual community attraction: Why people hang out online. Journal of Computer-Mediated Communication, 10(1), 00-00.

Saundage, D., & Lee, C. Y. (2011, January). Social commerce activities–a taxonomy. In ACIS 2011: Identifying the information systems discipline: Proceedings of the 22nd Australasian Conference on Information Systems. ACIS.

Turban, E., Whiteside, J., King, D., & Outland, J. (2017). Introduction to electronic commerce and social commerce. Springer.

Villeneuve, H., & O'Brien, W. (2020). Listen to the guests: Text-mining Airbnb reviews to explore indoor environmental quality. Building and Environment, 169, 106555.

Wirth, R., &Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39).

Zwass, V. (2010). Co-creation: Toward a taxonomy and an integrated research perspective. International Journal of Electronic Commerce, 15(1), 11-48.

Bibliographic information of this paper for citing:

Arzy, Amir, & Taghavifard, Mohammad Taghi, & Dehdashti Shahrokh, Zohreh, & Raeesi Vanani, Iman (2020). An Investigation on the User Behavior in Social Commerce Platforms: A Text Analytics Approach. *Journal of Information Technology Management*, 12(4), 180-199.