Original Scientific Paper

UDC: 366.622:077
338.488.2:640.4
doi: 10.5937/menhottur2001037R

# What do hotel guests really want? An analysis of online reviews using text mining

**Cvetanka Ristova Maglovska**[1*]

[1] Goce Delčev University of Štip, Faculty of Tourism and Business Logistics, North Macedonia

**Abstract:** Hotels offer a range of attribute-based services perceived to be wanted and gladly used by guests while staying at the hotel. That is, hotels at least think they have recognized the attributes of importance to their guests. Whether there is a desire for high-quality Wi-Fi, touchscreen technology, RFID or even tablet-controlled hotel room to satisfy the digital-savvy guests or small fridge, microwave and tea for families, hotels today find themselves into a position where online reviews represent one of the most valuable tools for getting insights into the factors that determine guests' experiences. By scraping the online reviews of 21 five-star hotels in North Macedonia on Booking.com, this paper investigates the attributes that are affecting guests' experience by analyzing the sets of online reviews using text mining. Research findings offer a more consistent understanding of the guest experience expressed in online reviews in terms of determining which amenities enhance guest satisfaction. The paper also illustrates how the methodological approach of text mining enables the use and visual interpretation of the data, and thus contributes to the studies in the field of hotel management.

**Keywords:** guest, hotel, online reviews, text mining
**JEL classification**: L83, Z32

# Šta gosti hotela stvarno žele? Analiza onlajn recenzije pomoću rudarenja teksta

**Sažetak:** Hoteli pružaju niz usluga zasnovanih na hotelskim atributima za koje pretpostavljaju da će ih gosti rado koristiti tokom boravka u hotelu. Odnosno, hoteli bar misle da su prepoznali atribute od značaja za njihove goste. Bez obzira na to da li postoji želja za visokokvalitetnom Wi-Fi tehnologijom, ekranom osetljivim na dodir, RFID ili hotelskom sobom koja je kontrolisana od strane tableta, ili pak za malim frižiderom, mikrotalasnom pećnicom i mogućnošću pripremanja čaja za porodicu, hoteli se danas nalaze u položaju gde onlajn recenzije predstavljaju jedan od najvrednijih alata za dobijanje uvida u faktore koji određuju iskustvo gostiju. Izdvajanjem onlajn recenzija 21 hotela sa pet zvezdica u Severnoj Makedoniji na Booking.com, ovaj rad istražuje atribute koji utiču na iskustvo gostiju analizirajući skupove onlajn pregleda primenom rudarenja teksta. Rezultati istraživanja nude dosledno razumevanje iskustva gosta izraženog u onlajn recenzijama u smislu utvrđivanja sadržaja koji povećavaju njegovo zadovoljstvo. Rad, takođe, prikazuje kako metodološki pristup rudarenje teksta omogućava korišćenje i vizuelno interpretiranje podataka, pa samim tim doprinosi studijama u oblasti hotelijerstva.

* cvetanka.ristova@ugd.edu.mk

**Ključne reči:** gost, hotel, onlajn recenzije, rudarenje teksta
**JEL klasifikacija**: L83, Z32

## 1. Introduction

Millions of guests are hosted by the hotel industry every day answering to all the expectations guests have when checking into the hotel. Hotels offer a range of attribute based services to their guests, but understanding what kind of services or attribute-based services align with the fulfillment of guests' expectations is the key to getting guests back, so it is not surprising that more and more hotels approach the use of advanced data analytics solution that will increase guest satisfaction (Shabani et al., 2017). The power of text mining has been proven in many industries, helping businesses save money, increase efficiency and make more informed decisions based on insightful activities. Because text mining helps raise the quality of service to a higher level, it is undoubtedly accepted by almost all sectors and businesses, and hospitality is no exception. Since the hotel industry is highly competitive (Jovanović, 2019), text mining finds its application following this fact. A key challenge for the hotel industry is that in an age of constant connectivity brought by digital technologies, guests have very high expectations and are increasingly looking for a personalized experience. It is actually the digital media that affects the modern lives of guests' (Vidaković & Vidaković, 2019), bringing them to a different pattern of consumption (Kuzmanović & Makajić-Nikolić, 2020) that craves for personalization i.e. specific desires for hotel attributes. If guests feel that they are not satisfied with the services that are offered, they always have many other options nearby. To maintain competitiveness and strengthen loyalty, hotels are taking certain steps such as using text mining, a helpful and valuable tool to analyze guests' expectations in order to increase the level of satisfaction. Due to the huge amount of data that guests create, text mining is the perfect partner for the hotel industry. Online reviews are considered as a form of data, representing a form of content created by the guests and shared on the digital platforms. So, when guests are happy, the Internet knows it. Similarly, when a guest is unhappy, hotels carry the burden of that resentment in their online reviews. But, by collecting and analyzing these data with text mining approach in real-time, hotels are offered an opportunity to enhance the guest experience through analyzing the attributes that cause it in the first place.

This research focuses on collecting the online reviews from the 21 five-star categorized hotels in North Macedonia at Booking.com and therefore present text mining as a tool that enables automatic scraping and extracting knowledge from unstructured text such as online reviews. The rationale for conducting this research is found in the effectiveness of text mining as a tool that has the ability to answer and deliver real-time decisions regarding the attributes that constitute the guests' experience from analyzing online reviews. The paper is structured as follows. First, the theoretical background discusses the existing literature on hotel attributes, guest-generated reviews on the Internet and text mining. Second, the research framework presents the usage of the text mining method with the details of stages of text analysis. Results and discussions are then elaborated through the document-term matrix, LDA and sentiment analysis from the online reviews. The paper concludes with a summary of the research findings, contribution, limitations and brief recommendations of text mining future usage in hotels operations.

## 2. Theoretical background

### 2.1. Guest experience and the importance of hotel attributes

While the basic desire of most guests at the hotel is the same - to have a place to sleep and keep their belongings - there are countless variations that exist above this basic need. Some

guests want to make use of many different hotel facilities and offerings, whether it is enjoying the pool, enjoying the spa or using a well-stocked gym. Yet, some guests will want to use more services, besides a bed in the room to sleep at night and store their clothes and other belongings in a safe place.

One of the most frequently researched topics in hospitality probably is identifying the attributes that actually contribute to hotel selection, booking, experience quality, guests' satisfaction and loyalty. Whereas, in academic literature, researches have shown that what hotel guests often want usually depends on their own individual specific needs, guests' needs vary from hotel to hotel, from destination to destination (Milićević et al., 2020). Such is the case with Israeli hotels for disabled guests, where Poria et al.'s (2011) research found that while staying at the hotel, guests faced difficulties with physical surroundings and customer service and interactions. A study by Rhee and Yang (2015) using combined analysis to compare guests' expectations with their actual experience in luxury hotels showed that the hotel classification influences upon the priority of hotel attributes desired by guests. Authors' findings ranked "rooms" as the most important guests' attribute, opposite to "location" which was ranked among the least important attributes of guests' experience and satisfaction.

Guests' experience and satisfaction in the past used to be researched and explored with guest surveys, especially comment cards (Lockyer, 2005). Today, with the rapid growth of digitalization, applications, and technology, guests are empowered to provide two-way information communication in the hotel industry as Internet users, thereby creating a huge amount of guest-generated content (Sigala, 2008) which ultimately provides the hotels a new method of understanding guests' expectations and experience adequately. Of all guest-generated content, online reviews are the ones characterised as electronic variations of the traditional written form of WOM (Filieri & McLeay, 2014), therefore being one of the most prominent tools (Chatterjee, 2001) to assess guests' experiences in hospitality.

Hotel managers have already found that guest-generated reviews improve their ability for well-informed decision-making within their management activities resulting in upgrading performance, maximization of hotel profitability, and guest loyalty. Nowadays, guest-generated reviews are used by hotels to minimize flaws and mistakes in the hotels' products and services and use hotel means effectively. Of course, it is clear that all these advantages in hotel operation are initially provided by the expressed positive feelings, i.e. guest satisfaction in online reviews and the described attributes of hotel services, referring to the experience. Academic researchers, who confirm this, state that only positive guest-generated reviews create a quality eWOM that serves to increase hotel online bookings (Torres et al., 2015) through amplifying the reputation of the hotel (Ye et al., 2009) and the guests' reliability in the hotel (Kim et al., 2009). Per Kim et al.'s (2015) study, hotel revenue is significantly impacted by the number of online reviews, and according to Tuominen's (2011) research, a strong positive connection exists among the hotel room occupancy rate and revenue per available room (RevPAR) and the number of online reviews generated by guests.

Obviously, products and services must meet guests' expectations (Lakićević & Sagić, 2019), therefore online reviews can be a huge benefit for hotels, but everything depends on how well they are managed. The key takeaway here is evident, and that is to invest in the cultivation of online reviews, because only through that process, will hotels be able to achieve the peak of guests' experience and satisfaction and therefore contribute to the overall hotels' operations. In what follows, text mining is presented as that key point able to examine and extract valuable accurate information from online reviews, i.e. attributes that form the guests' experience, because it is needless to say that, better performance of hotel operation activities listed above comes from fulfilled guests' experience and satisfaction, which in this case is derived from the hotel attributes can be traced in the guest-generated review.

### 2.2. Capturing the real value of online reviews with text mining

The challenge of gaining insights into guest-generated reviews is to extract valuable information and patterns from relatively large, highly unstructured (often messy) (McAfee & Brynjolfsson, 2012) text data written in natural language (authorized by human / guests). Manual scanning and analysis of such data is considered impractical due to the high computational load. Like the content itself, guest-generated reviews on the Internet are of high volume, variety and velocity and because of that could be considered as big data. However, how many data can be called "big" remains unknown (Mohanty, 2015). Since the volume actually refers to the number of online reviews generated on the Internet for a hotel in a given period, Liu et al.'s (2017) study shows the usage of 412,784 reviews generated by guests at 10,149 hotels on TripAdvisor in five Chinese cities and gives us insight about how "big" data could be. The authors used the advantage of the guest-generated reviews, segregating them per language group in order to provide practical knowledge about the drive behind guest satisfaction. Their study found that international tourists, who wrote online reviews in other languages (the research covers reviews written in English, French, Italian, German, Spanish, Japanese and Russian), significantly differed in their focus on the roles of various hotel attributes ("rooms", "location", "cleanliness", "service" and "value") in creating their maximum satisfaction rating of the hotel stay. As compared to international tourists, Chinese tourists at home had different room preferences relative to the other hotel attributes. Significant influences in the study were found regarding the interaction of hotel attributes "rooms" and "service" and between "value" and "service", meaning that if guests view a hotel deal as a good value for money, the demands they put on the service would be decreased, with the opposite impact at the other end of the value continuum.

Because of the unstructured form of natural language text and volume of guest-generated reviews on the Internet, text analysis, more precisely text mining and sentiment analysis play an essential part in capturing the real value from data. Text mining and sentiment analysis are considered the most suitable for various types of unstructured data in text form, where the content is legible and the meaning is obvious in the variety of various Internet sources, like review sites, social networks, forums, blogs, and so on. The objective of text mining is to exploit valuable information easily from a variety of text documents (He et al., 2013; Liu et al., 2011). Text mining's field deals with converting unstructured data into structured datasets, while reducing time and effort, but most of all it reduces data dimensionality into more manageable, meaningful and relevant data. Text mining focuses on finding and extracting hidden patterns, directions and connections, models, trends from unstructured data documents (He et al., 2015a; 2015b).

Recently, the approach of text mining on unstructured features of guest-generated reviews on the Internet, like text context, has been receiving growing academic attention (Goh et al., 2013). Actually, the research using text mining in hospitality is relatively recent. In hospitality, text mining frequently focuses on some common words listed in the online reviews, namely: location, room size, staff, breakfast, comfort, temperature, cleanliness and maintenance (Nickolas & Lee, 2017; O'Connor, 2010). For example, by analyzing and comparing 2,510 online reviews generated, Berezina et al. (2016) used text mining approach to investigate the basics of guests' satisfaction and dissatisfaction at a hotel. Authors' research further has found that satisfied guests who respond more frequently to the intangible elements as "staff" are more likely to recommend hotels to others than dissatisfied guests. Contrarily, dissatisfied guests rely more on tangible elements as "furniture" and "finances" during the hotel stay. In Barreda and Bilgihan's (2013) study of 17,357 guest-generated reviews from TripAdvisor.com it was found that "bedroom" and "bathroom" services, "location", "standards of service" and "cleanliness" were important issues for guests.

Since the Internet has brought information explosion of unstructured data, sentiment analysis (commonly known as opinion mining) has also become special application of text mining. Pang and Lee (2008) describe sentiment analysis as a computational systematic exploration and study of beliefs, emotions, thoughts, attitudes and subjectivity in a text. Further than this, sentiment analysis is considered to be a method for retrieving information and classifying data into subjective categories (positive, negative or neutral) or calculate the intensity of feelings (Pang & Lee, 2008; Thelwall et al., 2010). In academic research, Duan et al. (2013) operated the technique of sentiment analysis to extract 70,103 guest-generated reviews on the Internet from 1999-2011 of 86 hotels in Washington, while Geetha et al. (2017) concentrated on the sentiment orientation of guest-generated reviews on the Internet and found it had affected guest ratings. He et al. (2017) performed text mining, natural language preprocessing to clean the text documents and sentiment analysis for guest-generated reviews on the Internet and found that sentiment results from the title and content of guest-generated reviews were highly correlated with overall guests' ratings for hotels. The study of Qu et al. (2008) also suggests that most sentiment feelings derived from text reviews were significantly associated with overall guest assessment.

The above-mentioned research studies show how successful results and good outcomes from guest-generated reviews may come by leveraging text mining and sentiment analysis. The established relation between text analysis and guest-generated reviews in these studies further verifies that only by engaging this approach, hotel attributes mentioned in real-time online reviews that maximize the guests' experience can be determined.

## 3. Materials and methods

To obtain adequate and proper data, online reviews were collected from one of the biggest travel metasearch engines for reservation, Booking.com. Booking's reviews have been listed as more trustworthy because only guests that have previously booked through Booking.com can write a review, which eliminates the appearance of fake reviews. For the data to be compiled, Python web crawling was chosen, where Scrapy was used as a web crawler framework. As a research sample only five-star categorized hotels in North Macedonia were targeted at Booking.com, in order to understand what attributes cause guests' experience and satisfaction in luxury hotels. The crawler scraped 2,801 online reviews from 21 hotels on Booking.com from 8 cities in North Macedonia. As Kozinets (2010) states, there was now need for ethical clearance the "download of existing posts does not strictly qualify as human subjects research" (p. 151). Consent is only needed where there is contact or interference (Hookway, 2008).

For further analysis, the data was inputted into the R program. After the cleaning of non-English reviews, the data set was left with 2,222 online reviews written in the English language. Later, R program tools were used to process the data, and remove all stop words, punctuation marks and set all reviews into lowercase. Once the coding process was done, reviews were divided into words, resulting in 3,200 different words. When the word frequency was listed, a lot of adjective words were listed in the data, such as "wonderful", "amazing", "perfect" and others. Because this paper is focused on mining the attributes in the online reviews, some of these adjective words were removed, but some were left as "helpful" because they give us a closer look on how the attributes evoke the experience. Some other forms of linguistic entity were encoded in R into a "rudimentary" form reflecting the same significance. For example, for the word "restaurant", several forms as "restaurant", "restorant", "restoraunt" and "resturants" were noted and changed. Other unnecessary words include "look", "observation", "say", "round" and many others. This resulted with the extraction of 365 different words, connected to the guest experience.

## 4. Results and discussion

When all reviews and words were processed by R, the frequency of the words was calculated. This coding process was performed iteratively from high-frequency words to low-frequency ones. Table 1 displays the data from high-frequency to low-frequency words that were selected for statistical analysis with frequency of more than >30 words. Table 1 lists 70 relevant words associated to the guests' experience, used to describe the satisfaction rate alongside their total frequency. These words represent a broad variety of hotel and hotel stay related attributes relevant to the experience of hotel guests.

Table 1: Document frequency term-matrix of 70 words in online reviews

| Word | Frequency | Percentage | Word | Frequency | Percentage |
|---|---|---|---|---|---|
| room | 650 | 6.64% | price | 63 | 0.64% |
| hotel | 513 | 5.24% | bar | 62 | 0.63% |
| staff | 430 | 4.39% | children | 51 | 0.52% |
| breakfast | 398 | 4.07% | day | 51 | 0.52% |
| location | 331 | 3.38% | coffee | 49 | 0.50% |
| pool | 248 | 2.53% | cold | 49 | 0.50% |
| clean | 233 | 2.38% | lake | 47 | 0.48% |
| spa | 206 | 2.10% | reception | 47 | 0.48% |
| bathroom | 192 | 1.96% | expectation | 46 | 0.47% |
| comfortable | 187 | 1.91% | recommendation | 46 | 0.47% |
| center | 176 | 1.80% | booking | 45 | 0.46% |
| restaurant | 171 | 1.75% | roofbar | 45 | 0.46% |
| friendly | 159 | 1.62% | check-in | 44 | 0.45% |
| service | 150 | 1.53% | floor | 43 | 0.44% |
| food | 149 | 1.52% | free | 43 | 0.44% |
| helpful | 124 | 1.27% | polite | 43 | 0.44% |
| view | 116 | 1.18% | dining | 42 | 0.43% |
| bed | 109 | 1.11% | sauna | 42 | 0.43% |
| wellness | 91 | 0.93% | bad | 41 | 0.42% |
| place | 90 | 0.92% | hot | 38 | 0.39% |
| facilities | 88 | 0.90% | money | 38 | 0.39% |
| guests | 88 | 0.90% | noise | 38 | 0.39% |
| parking | 84 | 0.86% | superior | 38 | 0.39% |
| shower | 84 | 0.86% | warm | 36 | 0.37% |
| old | 82 | 0.84% | door | 36 | 0.37% |
| onebedroom | 81 | 0.83% | delicious | 35 | 0.36% |
| time | 81 | 0.83% | disgusting | 34 | 0.35% |
| area | 79 | 0.81% | relax | 33 | 0.34% |
| swimsuit | 78 | 0.80% | quality | 33 | 0.34% |
| quiet | 74 | 0.76% | balcony | 32 | 0.34% |
| renovation | 74 | 0.76% | broken | 32 | 0.33% |
| fitness | 71 | 0.74% | lights | 32 | 0.33% |
| star | 69 | 0.70% | air-condition | 31 | 0.33% |
| spacious | 68 | 0.69% | holiday | 31 | 0.32% |
| night | 63 | 0.64% | smell | 31 | 0.64% |

Source: Author's research

The frequency of these 70 words indicates that the first 18 words represent approximately half (46.39%), while the rest of the words account for less than 30%, (28.23%) of the total frequency of all words. As Ko (2018) states, the distribution can be described as one of the words with a fairly high frequency named as the "head" and words with low frequency (accurately in this research sample with an average percentage of less than 1 per word beginning from the 19th word) named "long tail". "Head" words, concentrate on basic products and services, as well as essential attributes, such as "room", "bed", "bathroom", "staff", "breakfast", "food", "service", "cleanliness", "location", "pool", "spa", "friendliness" and so on. The words from the "long tail" show the guests' other essential areas - the experience. Some of these words are practical and factual in nature, although others are guests' personalized evaluation of their stay in the hotel.

Afterwards, topic modelling was used trying to find similar topics across the guest-generated reviews on the Internet, and trying to group different words together, such that each topic will consist of words with similar meanings. The summary included in this paper was Latent Dirichlet Allocation (LDA) analysis with Gibbs sampling in R. LDA transforms each review into a vector of weights representing the intensity of each topic in the review, where a topic is a distribution of probabilities over the set of words used in the database reviews. This probabilistic model assigns the word a probability score of the most probable topic that it could be potentially belong to. Using these probabilities, twenty most likely words were ranged in each topic in order of how likely it was that each word belongs to that topic as shown is Table 2.

Table 2: Classifying words to a particular topic using LDA topic modelling

| **Services** | **Perception** | **Location** | **Value** | **Hybrid** |
|---|---|---|---|---|
| room | breakfast | building | restaurant | minibar |
| staff | roofbar | area | food | loud |
| breakfast | dining | spacious | view | renovation |
| bathroom | disgusting | apartment | time | delicious |
| restaurant | standards | new | amenities | air-condition |
| pool | dishes | nature | business | smell |
| spa | experience | walls | complementary | pleasant |
| facilities | tasteful | mountain | hospitality | furniture |
| price | cheap | trees | assistance | supermarket |
| reception | bread | motorcycle | categorization | luxury |
| recommendation | rich | monastery | room | local |
| free | healthy | horse | hotel | ambience |
| car | chocolate | mosquito | staff | vegetables |
| lobby | salad | room | location | connection |
| taxi | steak | hotel | service | suite |
| size | traditional | center | place | garden |
| sheets | omelet | view | facilities | complain |
| skybar | ham | place | price | payment |
| conference | cocoa | parking | free | attractions |
| netflix | staff | lake | booking | dust |

Source: Author's research

The words are in ascending order of phi-value. The higher the ranking, the more probable the word will belong to the topic. While performing LDA in R all 365 words were used in the analysis, in order to get a wider range of words. Topics were named according to the representation of words. Typical words for the five topics are:

1. Topic 1 (Services) – Usually associated words with the hotel's facilities include: "room", "breakfast", "bathroom", "pool" and "spa".
2. Topic 2 (Perception) – Perception words connected to the guests' stay include: "disgusting", "dishes", "cheap", "traditional" and "staff".
3. Topic 3 (Location) – Words associated with the location of the hotels include: "building", "area", "new", "nature", "center", "view" and "place".
4. Topic 4 (Value) – Words associated with guests' perceived value or money in the hotel include: "food", "time", "free", "complementary", "price" and "expectation".
5. Topic 5 (Hybrid) – Words that appear to consist two distinctive groups of words reflecting the very different experiences of hotel guests include: "vegetables", "supermarket", "delicious", "smell", "pleasant", "local", "luxury" and "dust".

Using NRC in R for sentiment analysis is immensely helpful when it comes to analyzing a text. In other terms, the polarity of the sentiment articulated within the spectrum that ranges from positive to negative is extracted and shown in Table 3.

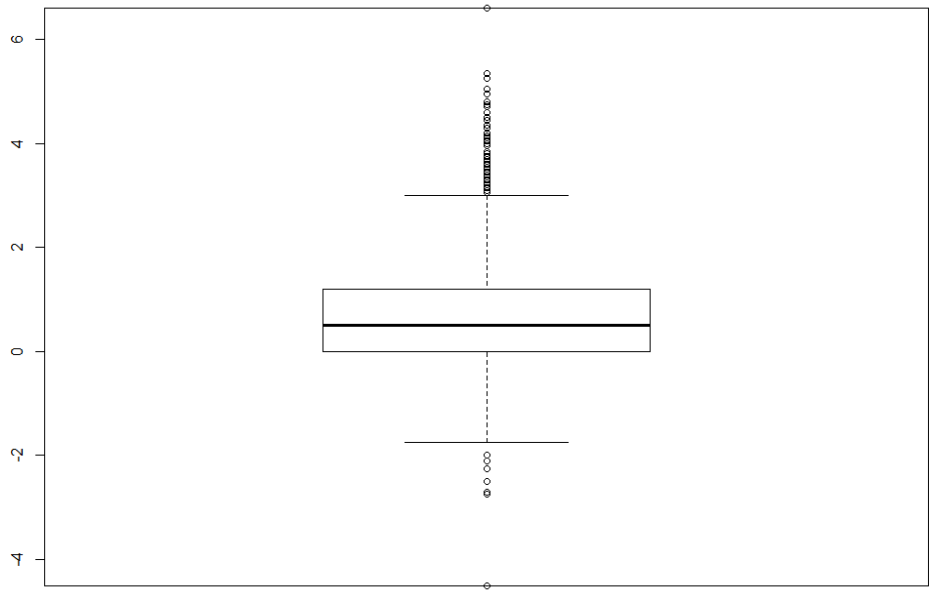
Table 3: Sentiment analysis of hotel online reviews

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -4.5000 | 0.0000 | 0.5000 | 0.6868 | 1.2000 | 6.6000 |

Source: Author's research

Our sentiment analysis has found that reviews tend to have slightly more positive content than negative content, but there are some extreme outliers with negative sentiment being -4.5, whereas positive sentiment was +6.6. As we see, these are quite far away from the mean and the median.

Figure 1: Boxplot of sentiment analysis of hotel online reviews




Source: Author's research

As a graphical representation of sentiment analysis, boxplot is used to visually show the distribution of numerical data and skewness through displaying the data. Seeing in Figure 1 that the median is +0.5, only a slight part of the online reviews tends to be positive. Since a

strong outlier is showed in both positive in negative sentiment, examples of the reviews with Syuzhet method in R will be shown in Table 4 to figure out which the most positive and negative review was and the whole paragraph.

Table 4: Examples of reviews with the highest and lowest sentiment

| Sentiment | Review |
|---|---|
| -4.5000 | The room was totally in a bad condition the furniture was a horrible, floor and the walls were full of dust, the jacuzzi in the room was broken and in a very bad and dirty condition. It was an improvisation of jacuzzi unfortunately. |
| 6.6000 | Beautiful hotel, beautifully decorated room, clean, excellent spa, super service at a high level, great choice of dining in the restaurant and finally an excellent choice for the end of the day in the roofbar on 5th floor. |

Source: Author's research

Given the statistics, guests' experience as seen in this research is, undoubtedly, an extraordinarily complex construct. In general, guests use the same attributes associated with hotel stays to assess their experience, although only the order of priority of those attributes may vary. The document-term matrix and topic modeling, identify the guest experience by representing which attributes guests consider as significant, important and contribute to their experience and satisfaction in a particular hotel. Therefore, these lists of words represent a "discreet" and "direct" presentation of the guest experience.

According to the research, guests' rank their experience as satisfactory or dissatisfactory in an online review from various perspectives. Guests complain most about dirty and outdated rooms, where the words "renovation", "old", "dirty" appear. However, data also noted, that guests were pleased with their hotel experiences for a variety of reasons, delicious food, friendly staff and location of the hotel in nature, mountains or even more particular they could be satisfied from products with rich, traditional attributes or dissatisfactory from products with disgusting attributes.

## 5. Conclusion

In these times, it is common for hotel guests to review a product or service online. In the online reviews, guests leave text content to publicly express their own personal thoughts and feelings, and briefly reveal their opinion about the hotel attributes. Critically, hotels are presented with a viable approach of the widespread applications of text analysis as text mining and sentiment analysis, allowing hotels to extract vast volume of accumulated text, and thus enabling them to effectively study and analyze the hotel attributes that form guests' experience. The research outcome shows that guests are prone to share even the tiniest detail of the hotel attribute and how it affected their experience in an online review. As expected, attributes as "room", "staff", "breakfast", "location", "bathroom", "restaurant" come among the top, but then again attributes as "shower", "balcony", "door" and "floor" appear, giving us more wide insight besides the already expected attributes mentioned above. Related attributes to the guests' stay at the hotel that also matter have been presented, where the findings reveal "friendly", "quiet", "polite", "delicious" and "disgusting" are important to the guests' experience.

In sum, the theoretical and practical implications of using text mining for scraping guest-generated reviews on the Internet are tremendously significant. Without text mining in hospitality, valuable guest information can be lost or ignored, which is a major shortcoming for a predominantly guest-oriented industry. Also, text mining can be used as a tool to create

vast profiles of information about guests, emphasizing the most important practical implication; enabling the hotel to deliver and create personalization or personalized experience.

Several limitations can be found in this paper. First, in this research sample, the document-term matrix presents a variety of hotel and hotel stay related attributes, where in future researches it is recommended to narrow them for better determination of the hotel attributes that contribute to the guests' experience. Second, LDA topic modeling shows some overlapping words, where again in future researches a small set of words might be used that would result in overlapping between topics. Lastly, NRC in R offers a possibility to further explore the sentiment in online reviews, beside positive and negative, but in another 8 emotion type, which by analyzing, can provide even better insights for hotels, like what attributes made guests feel joy, trust or anger, disgust.

## Acknowledgment

The research work by Ristova Maglovska, C. for this paper was done during an Erasmus+ Traineeship from February 1st, 2020 to May 1st, 2020 at the University of Kragujevac, Faculty of Hotel Management and Tourism in Vrnjačka Banja under the mentorship of assoc. prof. Darko Dimitrovski.

## References

1. Barreda, A., & Bilgihan, A. (2013). An analysis of user generated content for hotel experiences. *Journal of Hospitality and Tourism Technology*, *4*(3), 263–280. https://doi.org/10.1108/JHTT-01-2013-0001
2. Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, *25*(1), 1–24. https://doi.org/10.1080/19368623.2015.983631
3. Chatterjee, P. (2001). Online reviews: Do consumers use them? Online reviews: Do Consumers use them? *Advances in Consumer Research*, *28*, 129–133.
4. Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013). Mining online user-generated content: Using sentiment analysis technique to study hotel service quality. *46th Hawaii International Conference on In System Sciences (HICSS)* (pp. 3119–3128). Maui, Hawaii. https://doi.org/10.1109/HICSS.2013.400
5. Filieri, R., & McLeay, F. (2014). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information from online reviews. *Journal of Travel Research*, *53*(1), 44–57. https://doi.org/10.1177/0047287513481274
6. Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. *Tourism Management*, *61*, 43–54. https://doi.org/10.1016/j.tourman.2016.12.022
7. Goh, K. Y., Heng, C. S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, *24*(1), 88–107. https://doi.org/10.1287/isre.1120.0469
8. He, W., Tian, X., Tao, R., Zhang, W., Yan, G., & Akula, V. (2017). Application of social media analytics: A case of analyzing online hotel reviews. *Online Information Review*, *41*(7), 921–935. https://doi.org/10.1108/OIR-07-2016-0201
9. He, W., Tian, X., & Shen, J. (2015a). Examining Security Risks of Mobile Banking Applications through Blog Mining. *MAICS* (pp. 103-108). Greensboro, NC, USA: University of Dayton.

10. He, W., Wu, H., & Yan, G., & Akula, V., & Shen, J. (2015b). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(7), 801–812. https://doi.org/10.1016/j.im.2015.04.006
11. He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001
12. Hookway, N. (2008). Entering the blogosphere': Some strategies for using blogs in social research. *Qualitative Research*, *8*(1), 91–113. https://doi.org/10.1177/1468794107085298
13. Jovanović, S. (2019). Green hotels as a new trend in the function of sustainable development and competitiveness improvement. *Economics of Sustainable Development*, *3*(1), 1–7.
14. Kim, D. J., Ferrin, D. L., & Rao, H. R. (2009). Trust and satisfaction, two stepping stones for successful E-commerce relationships: A longitudinal exploration. *Information Systems Research*, *20*(2), 237–257.
15. Kim, W. G., Lim, H., & Brymer, R. A. (2015). The effectiveness of managing social media on hotel performance. *International Journal of Hospitality Management*, *44*, 165–171. https://doi.org/10.1016/j.ijhm.2014.10.014
16. Ko, H. C. (2018). Exploring big data applied in the hotel guest experience. *Open Access Library Journal*, *5*(10), e4877. https://doi.org/10.4236/oalib.1104877
17. Kozinets, R. V. (2010). *Netnography: Doing Ethnographic Research Online*. London, United Kingdom: Sage.
18. Kuzmanović, M., & Makajić-Nikolić, D. (2020). Heterogeneity of Serbian consumers' preferences for local wines: Discrete choice analysis. *Economics of Agriculture*, *67*(1), 37–54. https://doi.org/10.5937/ekoPolj2001037K
19. Lakićević, M., & Sagić, Z. (2019). Accommodation capacities and their utilization in the function of tourism development: Case of Ivanjica. *Ekonomika*, *65*(3), 77–88. https://doi.org/10.5937/ekonomika1903077L
20. Liu, B., Cao, S., & He, W. (2011). Distributed data mining for e-business. *Information Technology and Management*, *12*, 67–79. https://doi.org/10.1007/s10799-011-0091-8
21. Liu, Y., Teichert, T., Rossi, M., Li, H., & Hu, F. (2017). Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. *Tourism Management*, *59*, 554–563. https://doi.org/10.1016/j.tourman.2016.08.012
22. Lockyer, T. (2005). The perceived importance of price as one hotel selection dimension. *Tourism Management*, *26*, 529–537. https://doi.org/10.1016/j.tourman.2004.03.009
23. McAfee, A., & Brynjolfsson, E. (2012). Big data: The Management revolution. *Harvard Business Review*. Retrieved April 6, 2020 from https://hbr.org/2012/10/big-data-the-management-revolution.
24. Milićević, S., Đorđević, N., & Krejić, Z. (2020). Research on tourists' attitudes on the potential of Goč mountain for the development of eco-tourism. *Economics of Agriculture*, *67*(1), 223–238. https://doi.org/10.5937/ekoPolj2001223M
25. Mohanty, H. (2015). *Big data: An introduction*. *In Big Data: A Primer*. New Delhi, India: Springer.
26. Nicholas, W. C. K., & Lee, H. S. A. (2017). Voice of customers: Text analysis of hotel customer reviews (cleanliness, overall environment & value for money). *International Conference on Big Data Research* (pp. 104-111). Osaka, Japan: Association for Computing Machinery.
27. O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management*, *19*(7), 754–772. https://doi.org/10.1080/19368623.2010.508007

28. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011
29. Poria, Y., Reichel, A., & Brandt, Y. (2011). Dimensions of hotel experience of people with disabilities: an exploratory study. *International Journal of Contemporary Hospitality*, *23*(5), 571–591. https://doi.org/10.1108/09596111111143340
30. Qu, Z., Zhang, H., & Li, H. (2008). Determinants of online merchant rating: Content analysis of consumer comments about Yahoo merchants. *Decision Support Systems*, *46*(1), 440–449. https://doi.org/10.1016/j.dss.2008.08.004
31. Rhee, H. T., & Yang, S. B. (2015). Does hotel attribute importance differ by hotel? Focusing on hotel star-classifications and customers' overall ratings. *Computers in Human Behavior*, *50*, 576–587. https://doi.org/10.1016/j.chb.2015.02.069
32. Shabani, N., Munir, A., & Bose, A. (2017). Analysis of big data maturity stage in hospitality industry. Retrieved April 6, 2020 from https://arxiv.org/ftp/arxiv/papers/1709/1709.07387.pdf
33. Sigala, M. (2008). Web 2.0, social marketing strategies and distribution channels for city destinations: Enhancing the participatory role of travellers and exploiting their collective intelligence. In M. Gascó-Hernández, T. Torres-Coronas (Eds.), *Information communication technologies and city marketing: digital opportunities for cities around the world* (pp. 221-245). IDEA Publishing.
34. Thelwall, M., Buckley, K., & Paltoglou, G. (2010). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418. https://doi.org/10.1002/asi.21462
35. Torres, E. N., Singh, D., & Robertson-Ring, A. (2015). Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry. *International Journal of Hospitality Management*, *50*, 77–83. https://doi.org/10.1016/j.ijhm.2015.07.012
36. Tuominen, P. (2011). The influence of TripAdvisor consumer-generated travel reviews on hotel performance. *19th Annual Frontiers in Service Conference* (pp. 1–11). Columbus, Ohio, USA: University of Hertfordshire Business.
37. Vidaković, M., & Vidaković, D. (2019). Digital media, creativity, and marketing, within the scope of the contemporary instant culture. *The Annals of the Faculty of Economics in Subotica*, *41*, 131–144. https://doi.org/10.5937/AnEkSub1941131V
38. Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*, 180–182. https://doi.org/10.1016/j.ijhm.2008.06.011

Received: 13 April 2020; Sent for revision: 16 April 2020; Accepted: 4 May 2020