Information extraction and sentiment analysis of hotel reviews in Croatia

Šuman, Sabrina; Vignjević, Milorad; Car, Tomislav

Source / Izvornik: **Zbornik Veleučilišta u Rijeci, 2023, 11, 69 - 89**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.31784/zvr.11.1.5

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:191:267006

Original scientific paper
https://doi.org/10.31784/zvr.11.1.5
Received: 27. 12. 2022.
Accepted: 26. 2.. 2023.

INFORMATION EXTRACTION AND SENTIMENT ANALYSIS OF HOTEL REVIEWS IN CROATIA

Sabrina Šuman
PhD, Senior Lecturer, Polytechnic of Rijeka, Vukovarska 58, 51 000 Rijeka, Croatia; e-mail: ssuman@veleri.hr

Milorad Vignjević
Student, Polytechnic of Rijeka, Vukovarska 58, 51 000 Rijeka, Croatia; e-mail: milorad.vignjevic4@gmail.com

Tomislav Car
PhD, Assistant Professor, University of Rijeka, Faculty of Tourism and Hospitality Management,
Primorska 46, 51410 Opatija, Croatia; e-mail: tcar@fthm.hr

ABSTRACT

Today, the amount of data in and around the business system requires new ways of data collection and processing. Discovering sentiments from hotel reviews helps improve hotel services and overall online reputation, as potential guests largely consult existing hotel reviews before booking. Therefore, hotel reviews of Croatian hotels (categories three, four, and five stars) in tourist regions of Croatia were studied on the Booking.com platform for the years 2019 and 2021 (before and after the start of the pandemic COVID-19). Hotels on the Adriatic coast were selected in the cities that were mentioned by several sources as the most popular: Rovinj, Pula, Krk, Zadar, Šibenik, Split, Brač, Hvar, Makarska, and Dubrovnik. The reviews were divided into four groups according to the overall rating and further divided into positive and negative in each group. Therefore, the elements that were present in the positive and negative reviews of each of the four groups were identified. Using the text processing method, the most frequent words and expressions (unigrams and bigrams), separately for the 2019 and 2021 tourism seasons, that can be useful for hotel management in managing accommodation services and achieving competitive advantages were identified. In the second part of the work, a machine learning (ML) model was built over all the collected reviews, classifying the reviews into positive or negative. The results of applying three different ML algorithms with precision and recall performance are described in the Results and Discussion section.

***Key words**: hotel review, Booking.com, sentiment analysis, text processing, machine learning model*

1. INTRODUCTION

The tourism sector is of great importance to today's economy and will remain so in the coming decades (Kontogianni and Alepis, 2020). As hotels play an important role in the tourism sector, their performance is closely linked to the overall performance of tourism (Mucharreira et al., 2019). The hospitality industry is affected by the rapid growth of reviews and all types of user-generated content (UGC) on the Internet, and there is a need to implement various Big Data analytics methods to gain valuable insights (Mayer-Schoenberger and Cukier, 2013; Tsai et al., 2022).

(Liu, 2021) and (Onuiri et al., 2016) point out that smart tourism involves the use of ICT methods to fully leverage the vast amounts of data in the tourism industry for decision making and management. Incorporating AI (Artificial Intelligence) methods into Big Data analytics means that they can continuously learn and improve from all the input data analysed and predict customer behaviour.

UGC has a major impact on users' purchasing decisions - about 35% of travellers change their hotel decisions after reading relevant content on social media, while 53% say they would not book a hotel without reviews, and 87% say reviews increase confidence when choosing accommodation (Nicoli and Papadopoulou, 2017). Electronic Word of Mouth (eWOM) consumers write and they do not disappear immediately - other consumers can read these messages for a long time (Breazeale, 2009) and they become a reference point for buyers of goods or services. (Rita et al., 2022; Mou et al., 2022; Meng et al., 2022; Zenggang et al., 2022). The research findings of (Schuckert et al., 2015), who examined 50 articles on eWOM in hospitality and tourism, show that "online reviews seem to be a strategic tool that plays an important role in hospitality and tourism management, especially in promotion, online sales, and reputation management" (Martin-Fuentes, Mateu, and Fernandez, 2018). Online customer reviews (OCRs), as a particular form of eWOM, have a great impact on consumer decision making. OCRs are any positive, negative, or neutral feedback about a product, service, brand, or person provided and shared online (e.g., Booking.com, TUI.com, Facebook, Google reviews, etc.) by a past buyer (Hennig-Thurau et al., 2004; Filieri and Mariani, 2021)

The development of modern technologies in the tourism industry (recommendation systems, online reservations, dynamic pricing, and interactive platforms for evaluating services) have changed the way tourism products are consumed and the way consumers share their experiences and make decisions about choosing a new accommodation (Sanchez-Franco et al., 2019). It is also changing the way hotels should monitor and analyse online reviews to manage and improve service quality, as hotel reviews influence customers' booking intentions and reviewers' sentiment (Casaló et al., 2015; Rita et al., 2022; Mellinas et al., 2015)

Online platforms for selling tourism products generate a huge amount of data related to the service experience and form an online reputation of the hotel (Velázquez et al., 2015). The best way for hoteliers to build a good reputation and attract new customers is to manage the hotel's online reviews (Bridges, 2022). Therefore, it is crucial to identify the characteristics of customer satisfaction and dissatisfaction related to hotel quality.

The main objective of this research is to examine words and phrases appearing in positive and negative hotel reviews of a sample of Croatian hotels in ten locations on the Adriatic coast in 2019 and 2021 (before and after the pandemic COVID 19) on the Booking.com platform. Hotels were classified into four groups based on their overall rating: 7.0-8.0, 8.1-9.0, 9.1-9.4, 9.5-10.0.

In this information extraction process, several objectives are defined: to identify topics that appear in positive and negative guests' reviews, to investigate whether there are differences in guest perceptions before and after COVID 19 pandemics, and to find a sufficiently accurate ML model for polar sentiment. The following research questions were formulated:

RQ1) Is it possible to identify the main topics influencing positive and negative sentiment for four hotel rating categories (from 7-8, 8.1-9, 9.1-9.4, 9.5- 10) in the two years observed?

RQ2) Is there a difference in the topics of positive and negative reviews in 2019 and 2021? (Did the pandemic COVID -19 change the topics related to hotel service quality?),

RQ3) Is it possible to build a ML model to classify polar sentiment with acceptable performance (> =70% of precision and recall for each positive and negative class of ratings)?

After the introduction, the Related Work section reviews research in the field of hotel review processing and identifies areas related to this research and the innovations presented in this paper. The Methodology section describes the methodology used to collect reviews on the Booking.com platform, the data sample, and the reason for choosing this platform. It also describes the software tool that was used to create text analysis processes and ML models for 4 different categories of reviews. This is followed by the results and discussion describing the results of the text mining analysis for the reviews, which are divided into four categories by ratings for 2019 and 2021, before and after the COVID years. A ML model for classifying reviews as positive or negative in a sample of all reviews and for each of the above 4 categories is described, along with the performance of the classifier. In the conclusion, answers to the research questions are provided, limitations of this research are described, and directions for further research are given.

The contribution of this research aims at the service quality management in the hotel industry: a better understanding of guests' experiences and their perception of quality, as well as the possibility of using AI methods to obtain rapid analysis and valuable information about guests' feedback. In addition, the results of the research show the need for the future creation of a sentiment dictionary specifically for the field of tourism reviews.

2. RELATED WORK

Sentiment analysis provides information for understanding public opinion and analysing various tweets and reviews (Tul et al., 2017). The essence of the whole sentiment analysis is to classify the text and determine the contribution of different words for different classifications (Xu et al., 2019). There are many research papers on sentiment analysis in tourism sector, most of them are from 2018 and newer. Various authors have tried to discover word vector predictors of review polarity,

develop a reliable ML model for aspect-based sentiment classification, and provide some new insights into travellers' opinions and sentiment in the tourism industry.

Martin et al. (2018) investigated different Deep Learning techniques (based on Convolutional Neural Networks -CNN and Long short-term memory - LSTM) in the field of classification of online tourist comments (sentiment) from Booking.com and Tripadvisor. The LSTM recurrent neural network algorithm provided the most accurate results. Setiowati and Setyorini (2018) extracted service words and opinion words in their study and then identified sentiments from opinions on service quality indicators. They were then segmented by hotel department and function. Among k-Nearest Neighbour (KNN), Support Vector Machine (SVM), J48, and Naïve Bayes (NB), the rule-based method was used and achieved the highest precision, recall, and f-measure. To investigate the methods of measuring online reputation of hotels, (Pollak et al., 2018) applied a multifactorial analysis of online reputation (Google, Booking.com, Tripadvisor, and Facebook) and discovered a relationship between online reputation factors. The authors in (Mishra et al., 2019) performed text processing of hotel reviews by using TF-IDF and cosine similarity to extract similar values from the sentiment dataset. In their research, (de Brito et al., 2020) presented the development of SentimentALL, a sentiment analysis tool that extracts and analyses user comments from an online booking platform for travel services. The research of (Mostafa, 2020) proposed a Traveller Review Sentiment Classifier that analyses travellers' reviews about Egyptian hotels and provides a classification of hotel characteristics by sentiment. Among SVM, NB, and Decision Tree, NB had the highest accuracy. Stefko et al. (2020) assessed perceptions of service quality (polarity of sentiment) based on indicators such as location, staff rating, cleanliness, amenities, comfort, value/money, and Wi-Fi using regression analysis techniques. The results show that the cleanliness and amenities categories have the greatest impact on perceptions of a hotel's service quality. One of the research topics was also the sentiment of online consumers towards the environmental discussion based on hotel reviews in America and Europe, which increased over time (Mariani and Borghi, 2020). The authors (Oliveira Lima et al., 2021) compared hotel review classifiers using Latent Dirichlet Allocation (LDA), NB, logistic regression, SVM and LSTM which performed the best. The research of (Mehta et al., 2021) aims to evaluate customer satisfaction through sentiment analysis of customer reviews in the pandemic year 2020. The authors also conducted topic modelling to assess the most discussed topics by customers (12 most discussed topics and dissatisfaction with staff, service, room, cleanliness, slow booking, and hotel response to the pandemic). Sontayasara et al. (2021) created a ML sentiment analysis model using SVM with a classification accuracy of 71% for negative, positive, and neutral classes using Twitter data from the 2020 pandemic year. The changes that the COVID -19 pandemic brought to the hotel industry led to changes in guest perceptions of service quality attributes. The study by (Mušanović et al., 2021) provided a review of Facebook comments on hotel brand posts and applied sentiment analysis to identify and compare guest attitudes toward hotel staff, services, and products. The results showed that sentiment was more positive than negative and that there was no significant difference between the content and sentiments of the different hotel categories. The research identified words associated with positive and negative posts. (Peres and Paladini, 2022) examined the negative aspects affecting hotel service quality (a total of 13 aspects related to five hotel quality attributes) and found that room cleaning and check-in were the most negatively affected by the pandemic. Ghosal and Jain (2022)

used Word2Vec and extended families of Ordered Weighted Average (OWA) operators in their sentiment aggregation research. Their model includes explicit and implicit aspect segmentation for ratings, semantics for slang words, and location-based rating analysis. (Cendani et al., 2023) also used the LSTM model (with an attentional mechanism) for aspect-based sentiment analysis in their research.

The literature review showed that sentiment analysis of online tourism reviews is mainly based on finding a ML model that is accurate enough to classify polar sentiments regardless of the rating category. In line with the state of the art, this study also develops a ML model for all reviews, but also for each of the four review categories, to investigate the differences between positive and negative reviews in different hotel categories. It also explores the possibility that COVID -19 has changed the relevance of certain issues in relation to perceptions of hotel service quality. It is discussed whether it is necessary to create a specific sentiment dictionary for all types of sentiment analysis of OCR in the tourism sector.

3. METHODOLOGY

As mentioned in the introduction, it was necessary to collect reviews of hotels in the Republic of Croatia, and the Adriatic coast (and certain destinations) were chosen as hotel locations.

3. 1 Resources for the research data

The website "Touropia" (Best places to visit in croatia, 2022) lists Pula, Rovinj, Zadar, Split, Hvar and Dubrovnik as the top destinations for Croatian tourism in 2021. Taylor Herperger, in her article "15 Best Destinations in Croatia to Visit" (Herperger, 2022), adds the island of Brač and many other places like Makarska for their beauty and pleasant beaches. LonelyPlanet also mentioned a beautiful place in Croatia on its official website, namely the island of Krk, a place worth escaping other cities that have many more tourists. Šibenik is also mentioned in several sources. The following hotel locations were selected, whose reviews are considered in this article: Rovinj, Pula, Krk, Zadar, Šibenik, Split, Brač, Hvar, Makarska and Dubrovnik (Figure 1). All reviews are from the most popular website for hotels in the Republic of Croatia, Booking.com (https://www.booking.com/). Booking.com was selected based on the authenticity of the reviews and the rating methods described below. The Booking.com platform guarantees the authenticity and relevance of the reviews, as it allows reviews from people who have made a booking and completed a stay (at least one night in an accommodation). A review is then checked for inappropriate words and its authenticity is verified before publication. In addition, travellers can post positive and negative reviews separately on the Booking.com platform. This is important to determine customer satisfaction and dissatisfaction with the hotel's quality attributes (Booking.com, 2022; Peres and Paladini, 2022). Booking.com rates the property in six specific areas from 1-10: cleanliness, comfort, value for money, amenities, location, staff, and an optional open feedback. Starting in 2019, the overall rating is no longer the average of all six rating dimensions, but a new rating given by guests for the overall experience. This is due to the fact that guests may perceive other parameters not covered by the six specified (Booking.com, 2022).

Figure 1. Hotels' location on Adriatic coast

Source: Authors

3. 2 Research data

Reviews were collected for hotels in the Adriatic Sea before and after the Covid-19 outbreak (2019 and 2021). All reviews were from the most popular website for hotels in the Republic of Croatia, Booking.com (https://www.booking.com/). The ratings were divided into four groups (1st hotel group: 7.0-8.0 rating, 2nd hotel group: 8.1-9.0 rating, 3rd hotel group: 9.1-9.4 rating, and 4th hotel group 9.5 -10.0 rating), and for each of these groups, the ratings from 2019 and 2021 were considered separately. Table 1 shows the number of downloaded reviews by year and by one of the four groups: a total of 3117 reviews, 1600 positive reviews, and 1517 negative reviews (a smaller number of negative reviews from the hotel with the highest overall rating). Of the total 1,600 hotel facilities, 546 (34%) were 3-stars, 702 (44%) 4-stars, and 352 (22%) 5-stars.

Table 1. Number of reviews by year and hotel stars

Booking.com hotel guests' Ratings	Review Sentiment		Hotel stars		
	Positive	Negative	***	****	*****
7.0-8.0 (2019)	200	200	162	38	0
7.0-8.0 (2021)	200	200	156	44	0
8.1-9.0 (2019)	200	200	54	98	48

8.1-9.0 (2021)	200	200	54	98	48
9.1-9.4 (2019)	200	200	40	80	80
9.1-9.4 (2021)	200	200	40	80	80
9.5-10.0 (2019)	200	162	28	126	46
9.5-10.0 (2021)	200	155	12	138	50
SUM	**1600**	**1517**	**546**	**702**	**352**
TOTAL	**3117**			**1600**	

Source: Authors

Table 2 shows the number of reviews for each year and grouping from the selected 10 locations and the proportion of the number of reviews observed by location.

Table 2. Number of reviews from each location

Hotel Location	Booking.com overall guests' ratings									
	7.0-8.0 (2019)	7.0-8.0 (2021)	8.1-9.0 (2019)	8.1-9.0 (2021)	9.1-9.4 (2019)	9.1-9.4 (2021)	9.5-10.0 (2019)	9.5-10.0 (2021)	No. reviews	Share
Split	22	22	32	32	40	40	34	48	270	16,88%
Pula	36	40	38	38	20	20	28	28	248	15,50%
Hvar	22	10	22	22	20	20	36	36	188	11,75%
Dubrovnik	48	58	16	16	20	20	0	0	178	11,13%
Zadar	20	24	16	16	20	20	28	12	156	9,75%
Brač	20	12	16	16	20	20	22	22	148	9,25%
Krk	0	0	16	16	20	20	34	34	140	8,75%
Makarska	16	16	28	28	20	20	0	0	128	8,00%
Šibenik	16	18	0	0	0	0	18	20	72	4,50%
Rovinj	0	0	16	16	20	20	0	0	72	4,50%
SUM	200	200	200	200	200	200	200	200	1600	100,00%
TOTAL	**1600**									

Source: Authors

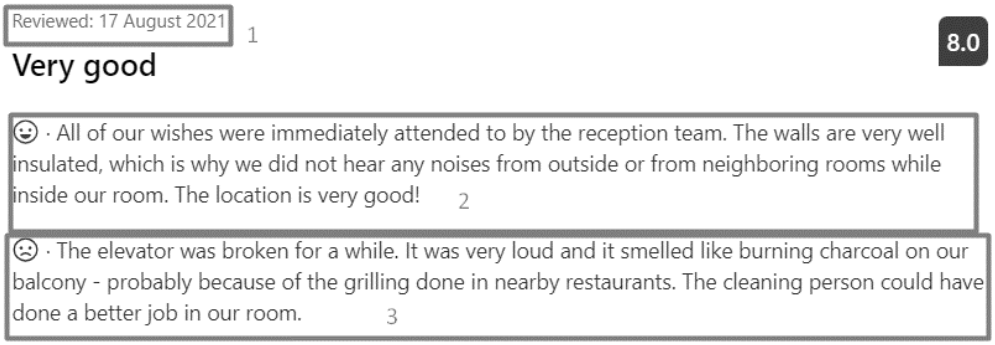
Once hotel groups were selected, hotel guest ratings were extracted for each group (separated into three-, four-, and five-star hotels). Figure 2 shows an example of a hotel from the first hotel group and framed relevant data. Label 1 represents the hotel's star rating, number 2 represents the hotel's location, number 3 represents the hotel's average rating, and number 4 represents the overall number of ratings.

Figure 2. Example of hotel data on Booking.com

Source: Authors

Figure 3 shows an example of a hotel rating used in data collection. Label number 1 indicates when the review was written, a very important aspect since only information for 2019 and 2021 was collected. Label number 2 contains a positive comment from the hotel, while label 3 contains a negative comment that had to be separated for later processing.

Figure 3. Example of review, positive and negative sentiment



Source: Authors

3. 3 Data analysis

After the data was collected, it was processed in RapidMiner software using various algorithms for text processing, sentiment analysis, and ML.

The RapidMiner data science platform was chosen for several reasons: It is open source, contains a large number of algorithms, the ability to add different packages, has a simple user interface, an intuitive way of working, and is regularly ranked as one of the best tools in its category (Wolff, 2020Hillier, 2022).

The first part of text processing was done using Data Operator's process documents (all reviews by category were structured in Excel spreadsheets and the relevant attributes were reviews in text form and sentiment - positive or negative), using tokenization operators (for word extraction) with mechanisms for cleaning and reducing word vectors by filtering stop words, eliminating words with less than 3 characters, setting lowercase letters and using Porter's stemming algorithm. The Term Frequency-Inverse Document Frequency method (TF-IDF) was used to obtain word vectors. In the Results and Discussion section, Tables 3, 4, 5 and 6. list the most frequently occurring words and phrases in positive and negative reviews for each hotel group in 2019 and 2021.

Then, using the RapidMiner operator Extract Sentiment and the Vader dictionary (Valence Aware Dictionary for Sentiment Reasoning), a sentiment analysis model was created to detect specific tokens for which the dictionary has an individual score (from - 4 to 0 are negative, 0 is neutral, and from 0 to 4 are positive). After the individual tokens, their scores are summarized and the overall sentiment score of a text (reviews in this case) is determined. The result of the sentiment analysis is described in the Results and Discussion section.

In the last part of the research, a ML model was built using operators in RapidMiner Deep Learning (DL), Gradient Boosted Trees (GBT) and Linear Support Vector Machine (LSVM), the results of which are described in the next section.

4. RESULTS AND DISCUSSION

The results obtained are presented below: The frequent words (unigrams) from positive and negative reviews, divided by hotel groups, the frequent bigrams from positive and negative reviews, divided by hotel groups, and ML models for polar sentiment extraction.

4. 1 Frequent unigrams

Tables 3 and 4 summarize the results by four groupings and years. The words with the highest frequencies were selected for display (the number of words in each category was not the same because we included only the most relevant frequencies). It should be mentioned that the application of stemming was also chosen when creating the word vector, so some words and phrases are in this form rather than in their original form as lexemes. Table 3 shows the most frequent words in the groups 7.0-8.0 and 8.1-9.0 (here there are the most hotels with 3 and 4 stars). No major variations were observed in the topics of positive and negative ratings in the years before and after the occurrence of Covid 19 - the occurrence of the word clean in the frequent words of negative ratings in 2021 after the pandemic was observed in both groups.

There are topics that appear mostly in **positive** reviews, such as*: staff, friendly, view, location, help, comfort, beach, love, beauty (stem)* and positive adjectives*: nice, good, great*.

Words/areas such as: *recept* (stem of reception), *bed, food, check* (check in and check out), *park* (parking), *restaurant, service, bathroom, old, book* (booking) are more common in **negative** reviews.

Areas that appear in both **positive and negative** reviews are *hotel, room, staff, breakfast, pool, clean* (where clean in negative reviews indicates a problem with cleanliness).

Table 3. Frequent words from group 7.0-8.0 and 8.1-9.0 ratings

Positive				**Negative**			
7.0-8.0 (2019)	**Freq.**	**7.0-8.0 (2021)**	**Freq.**	**7.0-8.0 (2019)**	**Freq.**	**7.0-8.0 (2021)**	**Freq.**
staff	89	**locat**	91	room	95	room	88
locat	67	**staff**	82	hotel	88	hotel	79
hotel	65	room	76	bed	66	breakfast	56
room	64	**good**	74	***recept***	65	old	43
beach	59	**nice**	70	breakfast	61	***check***	41
good	56	hotel	66	work	57	***recept***	39
view	52	**beach**	66	***check***	47	staff	38

breakfast	49	view	57	**old**	44	***bed***	35
great	41	breakfast	52	staff	44	peopl	32
clean	40	pool	51	book	38	clean	30
help	37	clean	51	night	29	star	26
nice	36	**great**	47	***food***	27	***food***	20
friendli	30	sea	43	locat	17	good	15
walk	26	**friendli**	38				
Positive				**Negative**			
8.1-9.0 (2019)	**Freq.**	**8.1-9.0 (2021)**	**Freq.**	**8.1-9.0 (2019)**	**Freq.**	**8.1-9.0 (2021)**	**Freq.**
room	98	hotel	105	room	133	room	116
breakfast	95	**great**	98	hotel	121	hotel	99
good	90	**staff**	96	pool	114	pool	91
locat	90	room	93	staff	88	breakfast	86
nice	90	breakfast	90	clean	82	beach	66
hotel	87	**locat**	90	beach	78	***bed***	59
great	84	**nice**	88	***bathroom***	69	clean	58
staff	81	**beach**	82	breakfast	65	view	47
pool	80	**good**	81	***bed***	63	peopl	45
beach	77	pool	77	towel	51	***bathroom***	44
clean	76	clean	72	good	51	area	43
love	75	food	60	***restaur***	44	***park***	43
walk	72	**love**	57	area	32	sea	43
area	66	**help**	55	close	31	***servic***	41
help	66	walk	47	***book***	28	time	22
view	48	**friendli**	43	make	16	***book***	21
food	45	**view**	42				
comfort	42	area	33				
friendli	39	**beauti**	31				

Source: Authors

In hotels with a higher overall rating, where the highest percentage of hotels with 4 and 5 stars are found, the word occurrence is similar to the previous two groups, except that the *pool* appears more often in **negative** reviews and the topics of *payment, price, noise and coffee* appear only in **negative** reviews. *Breakfast* is one of the most common topics in negative reviews across all groups and observed years, along with hotel and room. The word *excel* appears in **positive** reviews, where also appears more often word **comfort** (derived from excellent, excels...).

Table 4. Frequent words from group 9.1-9.4 and 9.5-10.0 ratings

Positive				Negative			
9.1-9.4 (2019)	**Freq.**	**9.1-9.4 (2021)**	**Freq.**	**9.1-9.4 (2019)**	**Freq.**	**9.1-9.4 (2021)**	**Freq.**
hotel	105	**staff**	117	room	98	hotel	121
staff	98	hotel	103	hotel	86	room	99
room	97	**locat**	101	breakfast	65	breakfast	87
locat	84	breakfast	96	pool	53	pool	79
breakfast	81	**great**	88	staff	53	beach	65
great	78	room	82	beach	45	**bed**	62
good	71	**friendli**	71	***restaur***	41	staff	61
friendli	64	clean	70	view	38	***restaur***	54
love	55	**good**	65	locat	37	**bathroom**	45
nice	52	**nice**	46	***book***	33	night	21
help	43	**help**	39	***pay***	31		
beach	42	**comfort**	35				
clean	41	pool	34				
excel	21	**view**	27				
Positive				**Negative**			
9.5-10.0 (2019)	**Freq.**	**9.5-10.0 (2021)**	**Freq.**	**9.5-10.0 (2019)**	**Freq.**	**9.5-10.0 (2021)**	**Freq.**
room	89	hotel	127	room	56	room	68
staff	86	**staff**	116	hotel	36	hotel	61
breakfast	82	room	103	breakfast	35	breakfast	59
hotel	82	breakfast	101	staff	33	***coffe***	42
locat	81	**great**	89	nice	28	night	39
good	75	**locat**	88	clean	22	staff	30
great	71	**good**	76	***restaur***	18	***restaur***	28
help	66	**help**	74	***servic***	18	peopl	26
friendli	63	**friendli**	73	***bed***	16	prefer	25
clean	62	**nice**	64	good	15	***price***	24
excel	58	clean	51	locat	15	star	19
comfort	44	**comfort**	31	**nois**	15		
nice	42	**excel**	25				

Source: Authors

4. 2 Frequent bigrams

After analysing the occurrence of individual words, an analysis of bigram searches and their frequency was performed for all hotel groups and the polarity of reviews. A combination of expressions was observed (mostly in the form of adjective_nouns), and it was determined which expressions occur most frequently in positive reviews, in negative reviews, or in both types of reviews. In this way, areas that are important to guests and that influence their satisfaction or are reasons for dissatisfaction are revealed.

Table 5. Frequent bigrams from group 7.0-8.0 and 8.1-9.0 ratings

Positive				**Negative**			
7.0-8.0 (2019)	**Freq.**	**7.0-8.0 (2021)**	**Freq.**	**7.0-8.0 (2019)**	**Freq.**	**7.0-8.0 (2021)**	**Freq.**
old_town	20	**sea_view**	14	valu_money	8	***dine_room***	8
friendli_staff	14	**friendli_help**	12	***air_condit***	6	***star_hotel***	8
beauti_view	12	**staff_friendli**	12	***book_doubl***	6	swim_pool	8
room_clean	12	**locat_good**	10	***doubl_bed***	6	***air_condit***	6
minut_walk	10	**nice_view**	10	***doubl_room***	6	***room_clean***	6
staff_friendli	10	**breakfast_good**	8	***hot_water***	6	***air_town***	4
staff_help	10	**good_locat**	8	***hotel_room***	6	amount_peopl	4
valu_money	10	**old_town**	8	***room_old***	6	***area_recept***	4
bus_stop	8	**room_clean**	8	***twin_bed***	6	ask_time	4
friendli_help	8	**adriat_sea**	6	activ_kid	4	***atm_min***	4
Positive				**Negative**			
8.1-9.0 (2019)	**Freq.**	**8.1-9.0 (2021)**	**Freq.**	**8.1-9.0 (2019)**	**Freq.**	**8.1-9.0 (2021)**	**Freq.**
great_locat	20	**staff_friendli**	18	***air_condit***	10	sea_view	10
pool_area	18	**breakfast_good**	16	room_bit	8	***coffe_machin***	8
locat_great	14	old_town	12	old_town	8	pool_area	8
staff_friendli	14	**great_locat**	12	room_clean	6	book_room	6
bed_comfort	12	pool_area	12	attent_detail	6	breakfast_buffet	6
locat_good	12	**staff_help**	12	chang_room	6	***park_place***	6
old_town	12	**breakfast_buffet**	10	outdoor_pool	6	peopl_hotel	6
breakfast_dinner	10	**minut_walk**	10	***room_balconi***	6	peopl_put	6
comfort_room	10	**access_beach**	8	staff_friendli	6	***put_towel***	6
buffet_good	8	**beach_great**	8	***bad_breakfast***	4	room_clean	4

Source: Authors

Positive reviews of all hotel groups (tables 5 and 6) are dominated by areas related to *friendly staff* and help, *bus_stop* and various positive adjectives along with *room, beach, buffet, view, location,*

and breakfast (nice, excellent, great, good, delicious, comfort...). Expressions that appear exclusively in **negative** reviews are *air_condition, reception, atm, air_town, hot_water, park_place* and negative adjectives with topics like *book, bed, room, balcony, breakfast, bathroom.*

The terms *beach_advertisement, citi_center, fridge, shower gel, wash_machine* were also found in the **negative** hotel reviews with ratings of 9.1-9.4 and 9.5-10.0.

The critical areas for both types of sentiments in the reviews turned out to be the following: *old_ town, breakfast, dinner, room_cleaning, room, sea_view, view, pool, booking, and value_money*.

Table 6. Frequent bigrams from group 9.1-9.4 and 9.5-10.0 ratings

Positive				**Negative**			
9.1-9.4 (2019)	**Freq.**	**9.1-9.4 (2021)**	**Freq.**	**9.1-9.4 (2019)**	**Freq.**	**9.1-9.4 (2021)**	**Freq.**
great_locat	23	**friendli_help**	26	sea_view	14	swim_pool	16
friendli_staff	18	**great_locat**	22	book_sea	13	**park_lot**	15
old_town	15	**good_breakfast**	18	**hotel_entranc**	13	**citi_center**	14
staff_help	15	old_town	18	old_town	13	sun_bed	14
breakfast_good	14	**staff_friendli**	16	pool_roof	11	star_hotel	12
good_breakfast	13	**staff_help**	15	star_hotel	9	**view_room**	11
great_hotel	8	**friendli_staff**	13	**view_room**	9	**access_beach**	9
help_friendli	8	**locat_great**	12	ac_sailor	8	**air_condit**	9
locat_good	8	**breakfast_good**	12	account_stai	6	**bathroom_door**	3
room_great	8	**comfort_room**	12	**adjac_room**	6	**beach_advertis**	3
Positive				**Negative**			
9.5-10.0 (2019)	**Freq.**	**9.5-10.0 (2021)**	**Freq.**	**9.5-10.0 (2019)**	**Freq.**	**9.5-10.0 (2021)**	**Freq.**
friendli_staff	24	**good_breakfast**	19	hotel_locat	13	star_hotel	14
excel_breakfast	21	**great_locat**	17	**bed_bit**	9	**coffe_breakfast**	9
friendli_help	21	**friendli_help**	17	**flight_stair**	9	hotel_riva	9
great_breakfast	20	**comfort_room**	15	**fridg_room**	8	**wash_machin**	9
room_clean	17	old_town	14	room_clean	6	person_prefer	8
sea_view	16	**sea_view**	13	**park_lot**	6	stai_bit	7
clean_good	14	**staff_friendli**	11	phone_call	6	room_bit	6
good_breakfast	12	swim_pool	11	room_nice	4	**year_old**	4
staff_friendli	12	**breakfast_great**	7	**shower_gel**	4		
breakfast_delici	7	**clean_comfort**	7				

Source: Authors

4. 3 Machine learning model for polar sentiment analysis

The research results can not only help in managing the quality of hotel services, but also serve as a basis for creating a sentiment dictionary that would include, in addition to standard words and their corresponding rating, these typical words for expressing sentiments in hotel ratings. For example, the word room would be paired with an adjective that refers to rooms and can be positive or negative. In this way, it would also be possible to create an aspect-based sentiment analysis that identifies sentiments related to an aspect, such as room. This is useful because when a sentiment analysis model was created using the Vader sentiment dictionary for all 3117 ratings, it was found that more than half of the negative ratings according to Vader were not negative. A look at the method of assigning the total score for each text unit (individual rating) shows that many negative semantics were not detected.

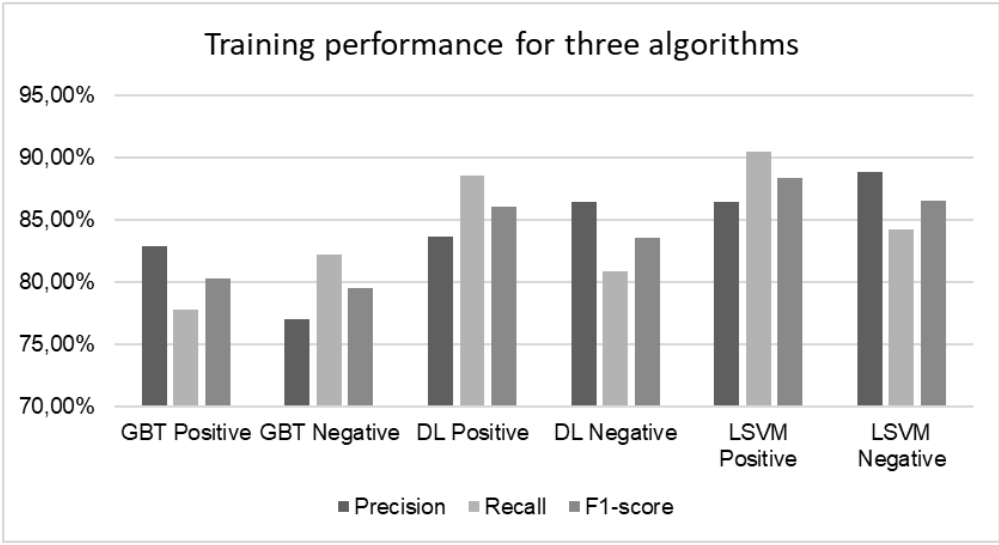
Since the existing sentiment dictionaries cannot detect the sentiment in a large number of reviews, the last part of the research was to build a ML model for detecting the sentiment of hotel reviews based on a specific ML algorithm and the number of reviews in the training phase. A training/testing partition with a ratio of 80:20 was created from a total of 3117 reviews from 2019 and 2021 (1600 positive and 1517 negative). The stratified sampling method was used (which ensures that the class distribution in the partitions is the same). Cross-validation was used for the training phase, which reduces the occurrence of overfitting by a factor of 10, and stratified sampling was also used for the folds. The following algorithms were investigated: Deep Learning (DL), Gradient Boosted Tree (GBT) and Linear Support Vector Machine (LSVM). The results are presented in Tables 7 and 8 and Figures 4 and 5. The results presented include the algorithms and parameters that provided the best results for the observed data partition.

Table 7. Performance results of training phase

	Class	Precision	Recall	F1-score
Training	GBT Positive	82.83%	77.78%	80.23%
	GBT Negative	76.99%	82.18%	79.50%
	DL Positive	83.60%	88.58%	86.02%
	DL Negative	86.48%	80.88%	83.59%
	LSVM Positive	86.40%	90.45%	88.38%
	LSVM Negative	88.87%	84.26%	86.50%

Source: Authors

Figure 4. Training phase performance of DL, GBT and LVSM algorithms



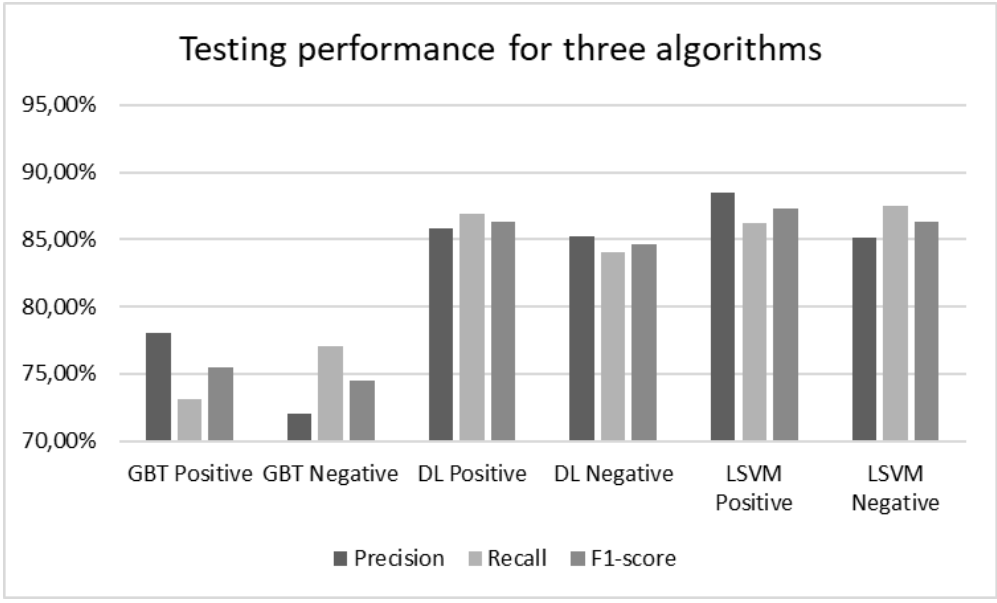
Source: Authors

Table 8. Performance results of testing phase

	Class	Precision	Recall	F1-score
Testing	GBT Positive	78.00%	73.12%	75.48%
	GBT Negative	72.08%	77.08%	74.50%
	DL Positive	85.80%	86.88%	86.34%
	DL Negative	85.21%	84.03%	84.62%
	LSVM Positive	88.46%	86.25%	87.34%
	LSVM Negative	85.14%	87.50%	86.30%

Source: Authors

Figure 5. Testing phase performance of DL, GBT and LVSM algorithms



Source: Authors

It can be seen that LSVM achieved the best overall performance of 87.51% in the training phase and 86.84% in the testing phase.

During research, other ML techniques were evaluated:

- the polar classification of each class of four evaluation groups - all performance parameters were below 75%.
- the classifier's ability to identify each rating group of the hotel based on the review - resulted in low performance (all performance parameters around 45% for both DL and GBT algorithms).

5. CONCLUSION

The adoption of modern technologies and Big Data analytics in tourism is necessary to monitor customer satisfaction, provide quality services and maintain competitiveness. OCR as a form of eWOM represents an important area of potentially valuable information and knowledge in the hospitality industry. The application of AI methods such as natural language processing and, in particular, sentiment analysis in the tourism sector makes it possible to gain some important insights that contribute to the management of the hospitality industry. There is a lot of research on text analytics, text mining, NLP, and sentiment analysis related to the hospitality industry, but with the development of ICT, there is still plenty of room for new insights. The hospitality industry is likely to be the most affected by coronavirus disease in 2019 (COVID -19). Therefore, the results

of the various sentiment analysis type studies should help hotel management to provide effective services to restore and maintain customer satisfaction.

This research was conducted on 3117 hotel reviews on the Croatian Adriatic coast in 2019 and 2021. Different text processing techniques and ML models were applied to answer three research questions RQs. The answers are as follows

RQ1) Is it possible to identify the main topics influencing positive and negative sentiment for four hotel rating categories (from 7-8, 8.1-9, 9.1-9.4,9.5- 10) in the two years observed?

Answer on RQ1): the topics that mainly appear in positive and negative reviews were identified, as well as the areas that appear in both types of reviews. The words and bigrams were used to determine which hotel services had the greatest impact on guest satisfaction and which were the main topics of negative sentiment and dissatisfaction. There were no significant differences among the four groups of hotels, except for some topics that occurred only in the group of hotels with the highest ratings.

RQ2) Is there a difference in the topics of positive and negative reviews in 2019 and 2021? (Did the pandemic COVID -19 change the topics related to hotel service quality?),

Answer on RQ2): there is no indication of a significant change in the topics appearing in the 2019 and 2021 reviews, with the exception of some new topics such as coffee, coffee maker and washing machine.

RQ3) Is it possible to build a ML model to classify polar sentiment with acceptable performance (> =70% of precision and recall for each positive and negative class of ratings)?

Answer on RQ3): for building classification models using ML, two main objectives were established: 1) to create a classifier that classifies a review as positive or negative, and 2) to create a classifier that can classify a review not only as negative or positive, but also as belonging to a particular hotel rating group. The results for 1) showed the performance of three ML algorithms, all of which achieved over 79% accuracy, with the LSVM algorithm achieving the best performance. The performance of the ML models for 2) was low, about 45% accuracy, for all observed algorithms.

The results of this study have highlighted the main strengths and weaknesses of the positive and negative scores and can be used to create action plans, eliminate problems, and maintain and improve the dimensions that are perceived as positive. The main limitations of this research are the relatively small number of assessments and the limitation to the Croatian Adriatic coast. Therefore, it is planned to expand the sample of ratings in the future and include more locations in different countries. Creating a sentiment dictionary for the tourism sector is also one of the goals of the next research, as well as exploring the extraction of aspect-based OCR semantics.

REFERENCES

Best places to visit in croatia (2022) Touropia.com. Available at: https://www.touropia.com/best-places-to-visit-in-croatia/.

Booking.com, P. H. (2022) Everything you need to know about guest reviews, Booking.com Partner Hub. Available at: https://partner.booking.com/en-gb/help/guest-reviews/general/everything-you-need-know-about-guest-reviews.

Breazeale, M. (2009) „FORUM - Word of Mouse - An Assessment of Electronic Word-of-Mouth Research", International Journal of Market Research, 51(3), pp. 1–19. doi: 10.1177/147078530905100307.

Bridges, J. (2022) 20 stats about online reviews that hoteliers need to know. Available at: https://www.reputationdefender.com/blog/online-reviews/20-stats-about-online-reviews-that-hoteliers-need-to-know.

Casaló, L. V et al. (2015) „Do online hotel rating schemes influence booking behaviors?", International Journal of Hospitality Management, 49, pp. 28–36. doi: https://doi.org/10.1016/j.ijhm.2015.05.005.

Cendani, L. M., Kusumaningrum, R. and Endah, S. N. (2023) „Aspect-Based Sentiment Analysis of Indonesian-Language Hotel Reviews Using Long Short-Term Memory with an Attention Mechanism", Lecture Notes on Data Engineering and Communications Technologies, 147, pp. 106–122. doi: 10.1007/978-3-031-15191-0_11.

de Brito, P. F., Tives, H. A. and Canedo, E. D. (2020) „Sentiment Analysis Tool in Website Comments", Advances in Intelligent Systems and Computing, 1134, pp. 709–716. doi: 10.1007/978-3-030-43020-7_97.

Filieri, R. and Mariani, M. (2021) The role of cultural values in consumers" evaluation of online review helpfulness: a big data approach, International Marketing Review. doi: 10.1108/IMR-07-2020-0172.

Ghosal, S. and Jain, A. (2022) „Weighted aspect based sentiment analysis using extended OWA operators and Word2Vec for tourism", Multimedia Tools and Applications. doi: 10.1007/s11042-022-13800-4.

Hennig-Thurau, T. et al. (2004) „Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?", Journal of Interactive Marketing. No longer published by Elsevier, 18(1), pp. 38–52. doi: 10.1002/DIR.10073.

Herperger, T. (2022) 15 Best Destinations in Croatia to Visit [in 2022], travellemming.com. Available at: https://travellemming.com/croatia-destinations/.

Hillier, W. (2022) What Are the Best Tools for Data Mining?, careerfoundry.com. Available at: https://careerfoundry.com/en/blog/data-analytics/best-data-mining-tools/.

Kontogianni, A. and Alepis, E. (2020) „Smart tourism: State of the art and literature review for the last six years", Array. Elsevier Ltd, 6(January), p. 100020. doi: 10.1016/j.array.2020.100020.

Liu, Y. (2021) „The Application of Big Data in the Intelligent Tourism Management Mode is Explored", Journal of Physics: Conference Series, 1881(3). doi: 10.1088/1742-6596/1881/3/032080.

Mariani, M. and Borghi, M. (2020) „Environmental discourse in hotel online reviews: a big data analysis", Journal of Sustainable Tourism, 29(5), pp. 829–848. doi: 10.1080/09669582.2020.1858303.

Martín, C. A. et al. (2018) „Using deep learning to predict sentiments: Case study in tourism", Complexity, 2018. doi: 10.1155/2018/7408431.

Martin-Fuentes, E., Mateu, C. and Fernandez, C. (2018) „Does verifying uses influence rankings? Analyzing booking.com and tripadvisor", Tourism Analysis, 23(1), pp. 1–15. doi: 10.3727/108354218X15143857349459.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Mehta, M. P., Kumar, G. and Ramkumar, M. (2021) „Customer expectations in the hotel industry during the COVID-19 pandemic: a global perspective using sentiment analysis", Tourism Recreation Research. doi: 10.1080/02508281.2021.1894692.

Mellinas, J. P., Martínez María-Dolores, S.-M. and Bernal García, J. J. (2015) „Booking.com: The unexpected scoring system", Tourism Management, 49, pp. 72–74. doi: 10.1016/j.tourman.2014.08.019.

Meng, F., Xiao, X. and Wang, J. (2022) „Rating the Crisis of Online Public Opinion Using a Multi-Level Index System", International Arab Journal of Information Technology, 19(4), pp. 597–608. doi: 10.34028/iajit/19/4/4.

Mishra, R. K., Urolagin, S. and Jothi, A. A. J. (2019) „A Sentiment analysis-based hotel recommendation using TF-IDF Approach", in Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019, pp. 811–815. doi: 10.1109/ICCIKE47802.2019.9004385.

Mostafa, L. (2020) „Machine Learning-Based Sentiment Analysis for Analyzing the Travelers Reviews on Egyptian Hotels", Advances in Intelligent Systems and Computing, 1153 AISC, pp. 405–413. doi: 10.1007/978-3-030-44289-7_38.

Mou, J. et al. (2022) „An effective hybrid collaborative algorithm for energy-efficient distributed permutation flow-shop inverse scheduling", Future Generation Computer Systems, 128, pp. 521–537. doi: https://doi.org/10.1016/j.future.2021.10.003.

Mucharreira, P. R. et al. (2019) „The relevance of tourism in financial sustainability of hotels", European Research on Management and Business Economics. AEDEM, 25(3), pp. 165–174. doi: 10.1016/j.iedeen.2019.07.002.

Mušanović, J., Dorčić, J., & Baldigara, T. (2021). Sentiment analysis of social media content in Croatian hotel industry . Zbornik Veleučilišta u Rijeci, 9(1), 37–57. https://doi.org/10.31784/zvr.9.1.3

Nicoli, N. and Papadopoulou, E. (2017) „TripAdvisor and reputation: a case study of the hotel industry in Cyprus", EuroMed Journal of Business. Emerald Publishing Limited, 12(3), pp. 316–334. doi: 10.1108/EMJB-11-2016-0031.

Oliveira Lima, T. D. et al. (2021) „A Big Data Experiment to Evaluate the Effectiveness of Traditional Machine Learning Techniques Against LSTM Neural Networks in the Hotels Clients Opinion Mining", in Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021, pp. 5199–5208. doi: 10.1109/BigData52589.2021.9671939.

Onuiri, E. E. et al. (2016) „Intelligent Tourism Management System", American Academic Scientific Research Journal for Engineering, Technology, and Sciences, 18(1), pp. 304–315. Available at: https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/1577.

Peres, C. K. and Paladini, E. P. (2022) „Quality Attributes of Hotel Services in Brazil and the Impacts of COVID-19 on Users" Perception", Sustainability (Switzerland), 14(6). doi: 10.3390/su14063454.

Pollak, F., Svetozarovova, N. and Malinak, B. (2018) „Multifactor analysis of online reputation as a tool for enhancing competitiveness of selected tourism entities", Global Business and Economics Review, 20(2), pp. 231–247. doi: 10.1504/GBER.2018.090074.

Rita, P. et al. (2022) „Impact of the rating system on sentiment and tone of voice: A Booking.com and TripAdvisor comparison study", International Journal of Hospitality Management. Elsevier Ltd, 104(May), p. 103245. doi: 10.1016/j.ijhm.2022.103245.

Sanchez-Franco, M. J., Cepeda-Carrion, G. and Roldán, J. L. (2019) „Understanding relationship quality in hospitality services", Internet Research. Emerald Publishing Limited, 29(3), pp. 478–503. doi: 10.1108/IntR-12-2017-0531.

Schuckert, M., Liu, X. and Law, R. (2015) „Hospitality and Tourism Online Reviews: Recent Trends and Future Directions", Journal of Travel & Tourism Marketing. Routledge, 32(5), pp. 608–621. doi: 10.1080/10548408.2014.933154.

Setiowati, Y. and Setyorini, F. (2018) „Service extraction and sentiment analysis to indicate hotel service quality in yogyakarta based on user opinion", in 2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018, pp. 427–432. doi: 10.1109/ISRITI.2018.8864269.

Sontayasara, T. et al. (2021) „Twitter sentiment analysis of bangkok tourism during covid-19 pandemic using support vector machine algorithm", Journal of Disaster Research, 16(1), pp. 24–30. doi: 10.20965/jdr.2021.p0024.

Stefko, R. et al. (2020) „Effect of service quality assessment on perception of TOP hotels in terms of sentiment polarity in the Visegrad group countries", Oeconomia Copernicana, 11(4), pp. 721–742. doi: 10.24136/OC.2020.029.

Tsai, Y. H., Lin, C. C. and Lee, M. H. (2022) „Analysis of Application Data Mining to Capture Consumer Review Data on Booking Websites", Mobile Information Systems, 2022. doi: 10.1155/2022/3062953.

Tul, Q. et al. (2017) „Sentiment Analysis Using Deep Learning Techniques: A Review", International Journal of Advanced Computer Science and Applications, 8(6). doi: 10.14569/ijacsa.2017.080657.

Velázquez, B. M., Blasco, M. F. and Gil Saura, I. (2015) „ICT adoption in hotels and electronic word-of-mouth", Academia Revista Latinoamericana de Administración. Emerald Group Publishing Limited, 28(2), pp. 227–250. doi: 10.1108/ARLA-10-2013-0164.

Wolff, R. (2020) 10 Best Data Mining Tools in 2022, MonkeyLearn. Available at: https://monkeylearn.com/blog/data-mining-tools/.

Xu, G. et al. (2019) „Sentiment analysis of comment texts based on BiLSTM", IEEE Access, 7, pp. 51522–51532. doi: 10.1109/ACCESS.2019.2909919.

Zenggang, X. et al. (2022) „A Service Pricing-based Two-Stage Incentive Algorithm for Socially Aware Networks", Journal of Signal Processing Systems, 94(11), pp. 1227–1242. doi: 10.1007/s11265-022-01768-1.

Izvorni znanstveni rad
https://doi.org/10.31784/zvr.11.1.5
Datum primitka rada: 27. 11. 2022.
Datum prihvaćanja rada: 26. 2. 2023.

EKSTRAKCIJA INFORMACIJA I ANALIZA SENTIMENTA HOTELSKIH RECENZIJA U HRVATSKOJ

Sabrina Šuman
Dr. sc., viša predavačica, Veleučilište u Rijeci, Vukovarska 58, 51 000 Rijeka, Hrvatska;
e-mail: ssuman@veleri.hr

Milorad Vignjević
Student, Veleučilište u Rijeci, Vukovarska 58, 51 000 Rijeka, Hrvatska;
e-mail: milorad.vignjevic4@gmail.com

Tomislav Car
Dr. sc., docent, Sveučilište u Rijeci, Fakultet za menadžment u turizmu i ugostiteljstvu,
Primorska 46, 51 410 Opatija, Hrvatska; *e-mail:* tcar@fthm.hr

SAŽETAK

U današnje vrijeme količina podatka koja se nalazi u poslovnom sustavu i oko njega zahtijeva nove načine prikupljanja i obrade podataka. Otkrivanje sentimenta iz hotelskih recenzija pridonosi poboljšanju hotelske usluge ali i ukupnoj online reputaciji budući da se potencijalni gosti prije rezervacije uvelike konzultiraju postojećim recenzijama smještaja. Na tragu toga, napravljeno je istraživanje nad hotelskim recenzijama hrvatskih hotela (kategorija tri, četiri i pet zvjezdica) u turističkim hrvatskim regijama sa platforme Booking.com, za godinu 2019 i 2021 (prije i poslije COVID 19 pandemije). Odabrani su hoteli sa Jadranske obale i to u gradovima koji su na više izvora odabrani kao najpopularniji: Rovinj, Pula, Krk, Zadar, Šibenik, Split, Brač, Hvar, Makarska te Dubrovnik Recenzije su grupirane u četiri grupe po ukupnom ratingu i dodatno podijeljene u svakoj grupi na pozitivne i negativne kako bi se identificirale stavke koje su prisutne u pozitivnim i negativnim recenzijama svake od četiri grupe. Metodom procesiranja teksta identificirane su najčešće riječi i izrazi (unigrami i bigrami) prisutni u spomenutim grupama recenzija, zasebno za 2019. i 2021. turističku sezonu, koje mogu poslužiti hotelskom menadžmentu kod upravljanja uslugama hotelskog smještaja i ostvarivanja konkurentske prednosti. U drugom dijelu rada, izrađen je model strojnog učenja nad svim prikupljenim recenzijama koji klasificira recenzije u pozitivne ili negativne. Rezultati primjene tri različita algoritma strojnog učenja sa performansama preciznosti i odziva opisani su u sekciji rezultati i diskusija.

Ključne riječi: *hotelska recenzija, Booking.com, analiza mišljenja, obrada teksta, model strojnog učenja*