Research Article
Journal of Intelligent Systems: Theory and Applications 6(2) (2023) 130-142
DOI: 10.38016/jista.1209415

Analyzing Big Social Data for Evaluating Environment-Friendly Tourism in Turkey

Mahmud Alrahhal[1*], Ferhat Bozkurt[2]

[1,2]Atatürk University, Department of Computer Engineering, Erzurum, Türkiye

alrahhal24@gmail.com, fbozkurt@atauni.edu.tr

Abstract
Tourism in Türkiye is fundamentally important for both the Turkish economy and travelers. Green tourism has gained increasing attention in the last few years. Analyzing big social data for evaluating environment-friendly tourism in Türkiye is important to gain an understanding of the factors impacting travelers' intention to echo-friendly hotels. To meet the goal of the study, the data was retrieved from the Tripadvisor website using a crawling technique. Machine learning techniques, particularly Latent Dirichlet Allocation (LDA), were utilized to discover satisfaction dimensions from the user-generated content. The k-means clustering approach was deployed for data segmentation. Finally, the online reviews classification model was trained and compared using Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The obtained results reveal several important dimensions that impact tourists' experience.

Keywords: LDA topic model, machine learning, text mining, segmentation, online customers' reviews.

1. Introduction

The tourism sector in Türkiye is furnishing increasingly. One of the most desired choices for travelers is going and unwinding in Türkiye. Many travelers prefer to spend their holiday in Türkiye due to the geographical position of Türkiye which is located in the heart of the world between Asia, and Europe, and it is close to Africa as well, along with its astonishing nature like long coasts, ancient structures, antiques, nice weather, and last but not least echo-friendly hotels. In 2019 alone, a total of 51.7 million travelers to Türkiye were recorded, with around 34.5 billion dollars in income, ranking the sixth worldwide in terms of the total number of tourists (Tuna & Başdal, 2021). These days, people are more willing to discover nature, since it improves their living standards as a consequence of relaxation, health, and taking advantage of environmental services (Prihayati & Veriasa, 2021). For several individuals, tourism is essential, and progressively earning importance. United Nations World Tourism Organization (UNWTO) conducted a survey revealing that near to 1.4 billion people traveled in 2019 (Streimikiene et al., 2021).

Adopting green services and products has increasingly become a center point in the growing businesses, with several organizations deploying environmental sustainability as a critical aspect of their main marketing policy (Chen et al., 2022). In the tourism sector, the green tourism concept has emerged and gained increasing attention (Filimonau et al., 2022; Yeşiltaş et al., 2022). Strong investigation of environmental issues, and looking for solutions to deal with such issues most of the time leads to a strong intention of travelers to perform echo-friendly actions to save the environment (Han et al., 2018). The initial expectations of the service given by a green hotel are primarily important for the customers. Measurement among customers' initial expectations and their actual experience of the product can describe customer satisfaction (Yu et al., 2022). Social big data analysis and online reviews are crucial to discover customers' expectations about many features located in environment-friendly hotels in Türkiye. Features extracted from online reviews will assist governments and decision-makers to know what are properties the customer interested in. Evaluating online reviews on green tourism, and environment-friendly hotel sites are substantially critical to enhancing both the Türkiye tourism sector and travelers' satisfaction. Although researchers have conducted many approaches and methods evaluating online reviews for tourism purposes, analyzing online reviews for environment-friendly hotels in Türkiye hasn't been investigated widely.

* Corresponding Author.
E-mail: alrahhal24@gmail.com

Recieved : 24 Nov 2022
Revision : 17 Feb 2023
Accepted : 31 Mar 2023

Online review extraction and mining have been the focus of researchers in natural language processing in the last few years (Afrizal et al., 2019). This is explained by the increasing popularity of online reviews among tourists, as 90% of them utilize these reviews to reach the travel decision and plan their trips (Godnov & Redek, 2016). Analyzing Social big data is fundamentally crucial to both customers and business owners. Big social data have a major impact on evaluating customer satisfaction with green hotels in Türkiye. Analysis of online reviews for green tourism in Türkiye will help decision-makers and hotel owners to define the features that the customers are interested in. Utilizing these shared online reviews which exist on hotel sites with a huge number of reviews is significant, as a customer is usually able to write any opinion about the hotel without any pressure from business owners or workers at that hotel. Unlike many products that are evaluated by amount or size, hotels are evaluated by experience (Zibarzani et al., 2022). Social big data consists of huge amounts of data, shared on several numbers of social media sites (Nilashi, Abumalloh, Almulihi, et al., 2021). Social big data analysis has been carried out in previous literature by utilizing different advanced approaches and methods to get reasonable assumptions and define market demands. Due to the increase of social big data and online reviews shared on the internet, Natural language Processing (NLP) has become indispensable not only for machine learning (ML) scientists but also for decision-makers in the market.

The main aim of this study is to explore the experience of tourists in environment-friendly hotels in Türkiye based on the content they post on TripAdvisor portal. To meet the goal of the study, we retrieved the data from the TripAdvisor portal, LDA was utilized to discover the dimensions of travelers' experiences, K-means algorithm was used to segment the customers according to their criteria ratings, LSTM, and GRU classification models was deployed and compared. To simplify the reading of this study, we present a list of abbreviations used in this study in Table 1.

Table 1. List of Abbreviations

Abbreviation	Full Term
ML	Machine Learning
NLP	Natural language Processing
LDA	Latent Dirichlet Allocation
UNWTO	United Nations World Tourism Organization
E-WOM	Electronic Word of Mouth
LDA	Latent Dirichlet Allocation
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit

2. Literature Review

2.1. Online reviews

Social media provides a substantial source for interaction, which can be utilized by many research fields like economy, trade, politics, and education (Bozkurt et al., 2019). The usefulness and reliability of online reviews stimulate the endorsement of reviews and the inclination of customers to trust online retailers (Shaheen et al., 2019). Customers' purchase decisions are heavily affected by online reviews (Huang et al., 2019). Customers' preferences located on online reviews for several sides of hotels not only affect customers' booking decisions but also help decision-makers to enhance the service quality of hotels continually (Bian et al., 2022). Online reviews have a considerable impact on determining the pricing strategy and increasing returns (Tian & Zhang, 2022). Online reviews represent a powerful type of communication and electronic word of mouth (E-WOM) where customers not only can spread their opinions but also can discuss their experiences, and it is a strong shape of hotel marketing (Nilashi, Minaei-Bidgoli, et al., 2021). Online reviews' reliability indicates the trustfulness of consumers to a specific review that is being read by the consumers (Wang et al., 2022). There is a powerful impact of online reviews and rating scores on product sales in a short timeframe (Ma et al., 2022). Research on online reviews for airline companies during the COVID-19 pandemic revealed extremely negative consequences due to several problems connected to refund strategies and procedures. Hence, online reviews provide decision-makers with a perception from customers' point of view of how airlines are able to deal with the serious effects of the COVID-19 pandemic (Rita et al., 2022).

2.2. Text mining, and LDA

The text analysis of online reviews ineluctably boosts the concept of "text mining" which indicates the procedure of extracting helpful and beneficial information from the unorganized text (Alzate et al., 2022). However, the textual nature of online reviews presents a complexity in analyzing and interpreting these social data (Alzate et al., 2022). There are diverse approaches for mining textual data, which have been deployed in the context of analyzing online reviews, including machine learning (Arulraj & Daisy, 2021) and lexicon-based methods (Xianghua et al., 2013). Each approach has its advantages and shortcomings. Machine learning approaches need an advanced level of experience in computational capabilities. As indicated by (Magoulas & Swoyer, 2020), finding skilled machine learning professionals by firms is not an easy task. As the analysis of texts requires specific requirements, Latent Dirichlet Allocation (LDA) was proposed by (Blei et al., 2003) to inspect the topics of textual data and to examine the level of competitiveness between the products. LDA adopts the concept of a "bag of words" to reflect the text as a combination of topics with multinomial dissemination of terms. The document entails topics, each document has topics with its own share and terms' distribution. As an unsupervised learning approach, it can locate the topics to reflect the wisdom of the crowds. Supervised approaches entail learning data that have a target. In context of textual data approaches Including LSTM, and GRU are used.

Recurrent Neural Network (RNN) models are widely utilized in sequential data modeling, including natural language, image/video, captioning, and prediction (Chimmula & Zhang, 2020; Khaldi et al., 2023). LSTM is Long Short-Term Memory Network model which is an extension of RNN (Hochreiter & Schmidhuber, 1997). Since gradient vanishing problem affect the RNN operation when dealing with longer sequence models, LSTM introduces memory cells consisting of different types of gate units, including "output gate", "input gate" and "gate forget" (Liang & Niu, 2022). As a different alternative of LSTM, gated recurrent unit (GRU) is analogous to LSTM in terms of performance, but its computational complexity is lower (Jung et al., 2018). GRU is a simpler, popular, and variant of LSTM and uses the same gate control mechanism as LSTM (Zhao et al., 2017).

2.3. Green tourism

Green hotels have been an interestingly important field of research in recent years, scholars have taken the topic into consideration growingly and increased publications related to the topic have been noticed (Acampora et al., 2022). For customers from several nationalities, there is a positive effect of hotels adopting eco-friendly practices on customers' satisfaction and customers' return inclination to environment-friendly hotels (Berezan et al., 2013). Environment-friendly hotels have a serious impact on customers' satisfaction and return intention (Merli et al., 2019). The definition of green hotels can be described as eco-friendly estates that apply eco-friendly behaviors like water saving, energy saving, and recycling to protect the globe that we inhabit (Association, 2008). Launched by TripAdvisor in 2013, the GreenLeaders program aims to encourage the adoption of green practices in US hotels (Chen et al., 2022). The research on this topic has integrated the tourism field along with sustainability aspects and gained increasing attention (Filimonau et al., 2022). Starting from 2016, the relationship between echo-friendly tourism and customer behaviors has been explored in more than 120 studies (Chen et al., 2022). Environmental issues represented by water and energy consumption, carbon emissions, and waste treatments have gained the attention of policymakers and induced them to focus on the production of green-friendly products and services (Verma et al., 2019)

3. Materials and Methods

3.1. Topic modeling (LDA)

Topic modeling technique utilizes statistical approaches to inspect unstructured texts and investigate the themes from them. This can be achieved through structuring the text within a number of topics that reflect the content of the text using Latent Dirichlet Allocation (LDA). In the LDA technique, the number of topics should be determined, indicating that 20 topics have provided the best performance by several studies (Williams & Betak, 2018). LDA has been deployed in text mining studies to explore online reviews and incorporate customers' perceptions in several areas of research such as marketing (Huang et al., 2022), online education (Wei & Taecharungroj, 2022), and accommodation business (Sim et al., 2021). Figure 1 presents the generative model of the LDA.

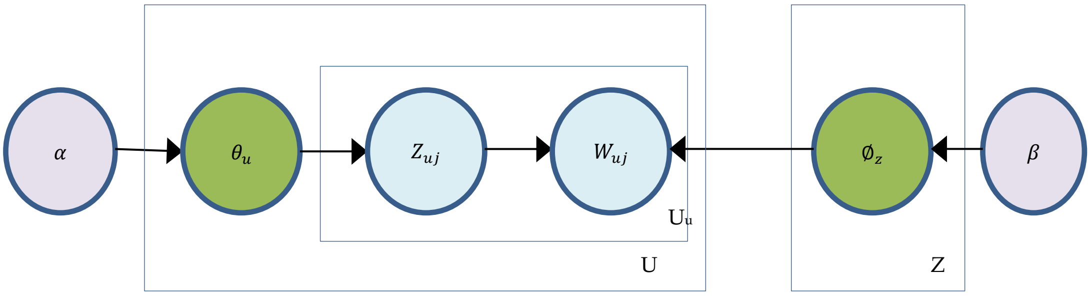


Figure 1. LDA Generative Procedure

Description of Figure 1.

1. For each topic $z \in Z$
   - Draw a multinomial distribution $\emptyset_z \sim Dir(\vec{\beta})$.
2. For every user $u \in U$,
   - Draw a multinomial distribution$\theta_u \sim Dir(\vec{\propto})$.
   - For every Word $w \in D_u$,

(a) Draw a topic $z \sim Multinomial\left(\underset{\theta_u}{\rightarrow}\right)$.

(b) Draw a word $w \sim Multinomial\left(\underset{\emptyset_z}{\rightarrow}\right)$.

3.2. Clustering Approach(K-means)

Following the topic modeling approach, a clustering technique was deployed to separate the user-generated content into several segments. Clustering approaches were utilized in several researches related to user-generated content analysis (Nilashi, Abumalloh, Alghamdi, et al., 2021; Nilashi et al., 2022). The k-means clustering approach was deployed as an iterative clustering technique that finds the optimal cluster center through several iterations (E. Zhang et al., 2022) and as an unsupervised technique. The deployed k-means method is illustrated in algorithm 1. In order to separate the n data into specific groups, the K-means algorithm finds the mean distance between data points. As presented in algorithm 1, the k-means clustering approach repeats the calculation of the distances between data points and assigns centroids to the specified clusters according to the updated distances until convergence.

Algorithm 1: iterative K-means clustering

```
Input    k: the number of clusters, X: A dataset with n data points
          Randomly initialize k centroids

Output   Set of centroids ($\mu_z$)
         Repeat
              Assignment of each data point to its closest centroids.
              Update the cluster centers ($\mu_z$)
         Until convergence
         Return ($\mu_z$)
```

3.3. Classification Model

Finally, In order to classify the reviews and predict the new ones in terms of being negative or positive we deployed and compared LSTM, and GRU machine learning methods. RNN is gaining increasing importance in natural language processing and text classification. The simplest RNN cell is ELMAN which is illustrated in Figure 2, which contains only one hidden layer.

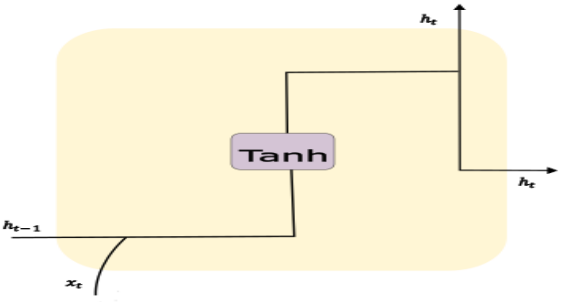


Figure 2. ELMAN Cell (Khaldi et al., 2023)

The output from the hidden layer in the RNN is also used as the input for the next value input along with the input value (Chen et al., 2018). In this way, RNN contains sequence dependency definition, for example, output (h_t) carries a dependency ratio to previous outputs as shown in Figure 2. Therefore, RNN is a successful recurrent neural network in predicting the next value (Wu & Noels, 2022).

The LSTM model was proposed by (Hochreiter & Schmidhuber, 1996) to solve the problem of gradient vanishing in RNN. LSTM offers memory cells consisting of several types of gate units, including “forget gate”, “gate gate”, and “exit gate” in each recurrent body. As shown in Equations (2.1)-(2.4), the LSTM unit adds input gate i, forgotten gate f, memory unit c, and output gate based on RNN, which significantly enhances the long sequence process performance (Wu et al., 2022). The operation of the LSTM unit is expressed by Equations (2.1)-(2.5).

$$i_t = \sigma(W_i \times [h_{t-1}] + b_i) \tag{1}$$

$$f_t = \sigma(W_f \times [h_{t-1}] + b_f) \tag{2}$$

$$c_t = f_t \times c_{t-1} + i_t \times tanh(W_c \times [h_{t-1}, x_t] + b_g) \tag{3}$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = o_t \times \tanh\,(c_t) \tag{5}$$

At time t i_t, f_t, f_t, c_t, o_t represent the input port, forget port, memory unit, and output port, respectively. H (hidden layer) represents the hidden layer. The x, w, b, and c are represented as input, weight, deviation, and cell respectively. (σ) represents the sigmoid function. An example of the LSTM unit structure was provided in figure 3.

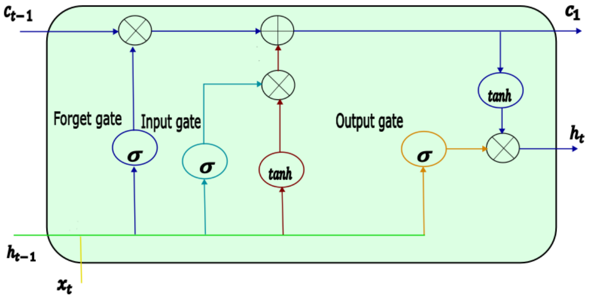


Figure 3. LSTM Structure (Wu et al., 2022)

GRU is similar to LSTM in terms of performance, however, its computational complexity is lower than LSTM and removes cell state, and uses a hidden state to transmit information (Jung et al., 2018). Along with solving the GRU Gradient vanishing problem, it combines the forget gate and input gate in LSTM into the update gate. The GRU consists of two gates, an update gate (update gate z_t) and a reset gate (reset gate r_t). In Figure 4, the structure of the GRU is provided. The operation of the GRU unit is expressed by Equations (2.1)-(2.4).

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{6}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{7}$$

$$h'_t = tanh(W x_t + r_t \odot U h_{t-1}) \tag{8}$$

$$h_t = z_t \odot U h_{t-1} + (1 - z_t) \odot h'_t \tag{9}$$

Where time is referred by t, x_t and h_t are input vectors. The weight matrices (W_z, U_z),(W_r, U_r),$(W_{h'}, U_{h'})$ represent the weights for the reset gate, update gate, and candidate latent state (h'), respectively. Σ represents the sigmoid function, ʘ represents the Hadamard product, and tanh represents the hyperbolic tangent function. The structure of GRU was provided in figure 4.

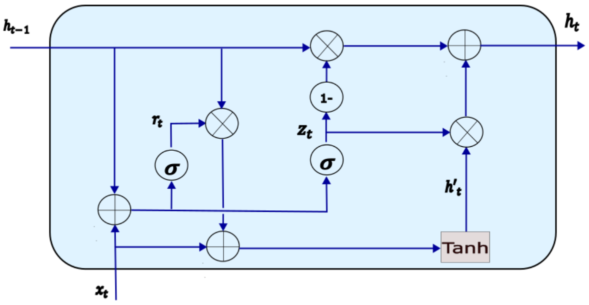
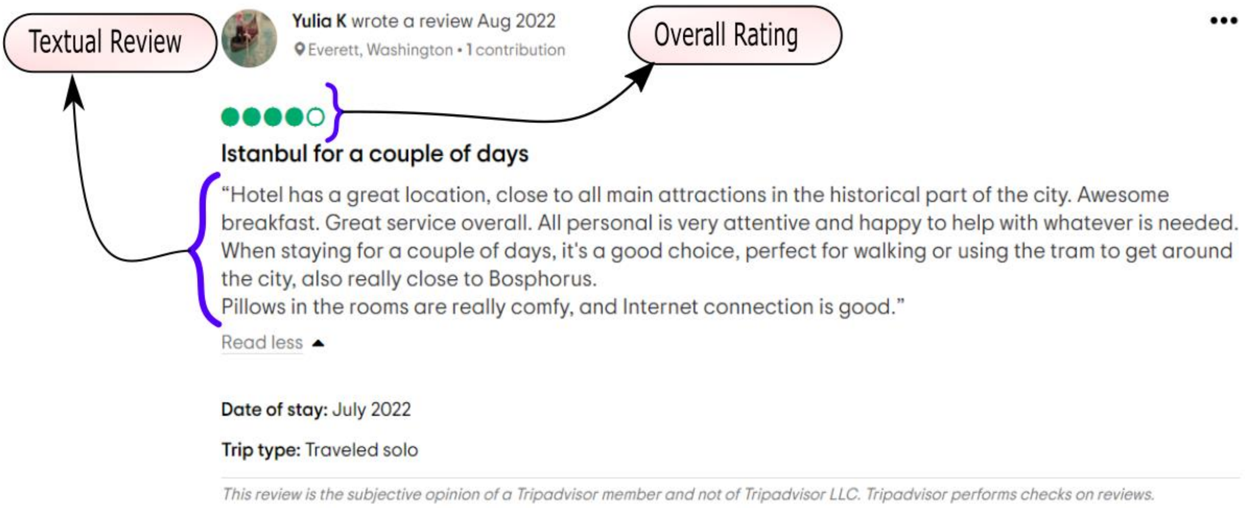


Figure 4. GRU Structure (C. Zhang et al., 2022)

3.4. Data Collection and Preprocessing

TripAdvisor was utilized to obtain the dataset for this research. Customers' online reviews were gathered from different eco-friendly hotel websites located in Türkiye; which are presented on the TripAdvisor platform. The TripAdvisor GreenLeaders program regards the behaviors of hotels towards green practices and ranks them according to 4 levels: Bronze, Silver, Gold, or Platinum; which are displayed notably on the estate's listing on the TripAdvisor site. Properties which demonstrate more green actions are able to get higher TripAdvisor GreenLeaders levels (UNEP, 2013). The crawling technique was employed to crawl TripAdvisor hotel sites using their URLs. Selenium library was utilized to crawl the online reviews from different green hotels located on the TripAdvisor website. Webdriver was imported from the Selenium library and google chrome was utilized for the crawling operation. Reviews located on TripAdvisor hotel sites are distributed with around 10 reviews per page. In the crawling technique, we deployed a loop to navigate through these pages and get the body of the reviews by their data-reviewid XPath then get reviews by their XPath. The looping operation utilized Selenium and for each iteration, to navigate to the other pages the next button of the page was clicked and the next URL was given to the crawler. The operation continues throughout the pages until reaching the last page of the green hotel located on TripAdvisor. Figure 5, illustrates an example of the text-based reviews collected by means of the crawler. The crawler was built to collect customers' online reviews related to the hotels that we aimed to investigate. We gathered 17314 online reviews from different hotels located in Türkiye. The collected reviews' language is English. Collected data was cleaned from useless words or sentences like emails and new line symbols. In addition, gathered data was checked in terms of the existence of null values. We avoided encountering unfamiliar vocables and texts in the results by cleaning the collected data. Criteria ratings have been collected by the crawler and missing values have been filled using the mean of the column to which the data point belongs. Figure 6. illustrates the criteria ratings generated by the users. The research method followed in this study is presented in Figure 7.

In the customer review classification stage, we noticed that the customers' reviews with 4 overall ratings and higher were positive and the customers' reviews with 3 overall ratings and lower were negative. The dataset was consisting of 951 negatives and 16363 positive reviews. In order to train the model with a balanced dataset we collected more data with negative reviews and eventually, we built a new balanced dataset for the classification model with an overall of 6611 reviews consisting of 3305 positive reviews and 3306 negative reviews. The dataset was separated into train and test, 80% of the dataset was allocated for train and 20% for test. To train the ML model, the customers' reviews were converted to numerical values using Tokenizer API from TensorFlow Keras. Eventually, the sentences are represented by a sequence of numbers using texts_to_sequences from the Tokenizer object.



Yulia K wrote a review Aug 2022
Everett, Washington • 1 contribution

●●●●○

Istanbul for a couple of days

"Hotel has a great location, close to all main attractions in the historical part of the city. Awesome breakfast. Great service overall. All personal is very attentive and happy to help with whatever is needed. When staying for a couple of days, it's a good choice, perfect for walking or using the tram to get around the city, also really close to Bosphorus.
Pillows in the rooms are really comfy, and Internet connection is good."

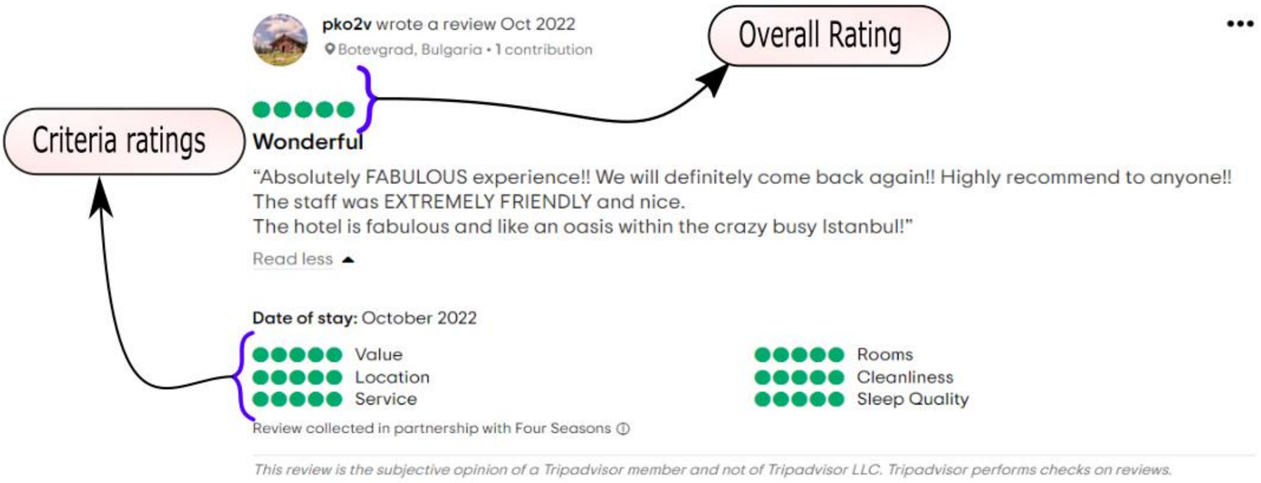
Read less ▲

Date of stay: July 2022

Trip type: Traveled solo

This review is the subjective opinion of a Tripadvisor member and not of Tripadvisor LLC. Tripadvisor performs checks on reviews.

Figure 5. Textual Review



pko2v wrote a review Oct 2022
Botevgrad, Bulgaria • 1 contribution

●●●●●

Wonderful

"Absolutely FABULOUS experience!! We will definitely come back again!! Highly recommend to anyone!!
The staff was EXTREMELY FRIENDLY and nice.
The hotel is fabulous and like an oasis within the crazy busy Istanbul!"

Read less ▲

Date of stay: October 2022

●●●●●	Value	●●●●●	Rooms
●●●●●	Location	●●●●●	Cleanliness
●●●●●	Service	●●●●●	Sleep Quality

Review collected in partnership with Four Seasons ⓘ

This review is the subjective opinion of a Tripadvisor member and not of Tripadvisor LLC. Tripadvisor performs checks on reviews.

Figure 6. Criteria Ratings

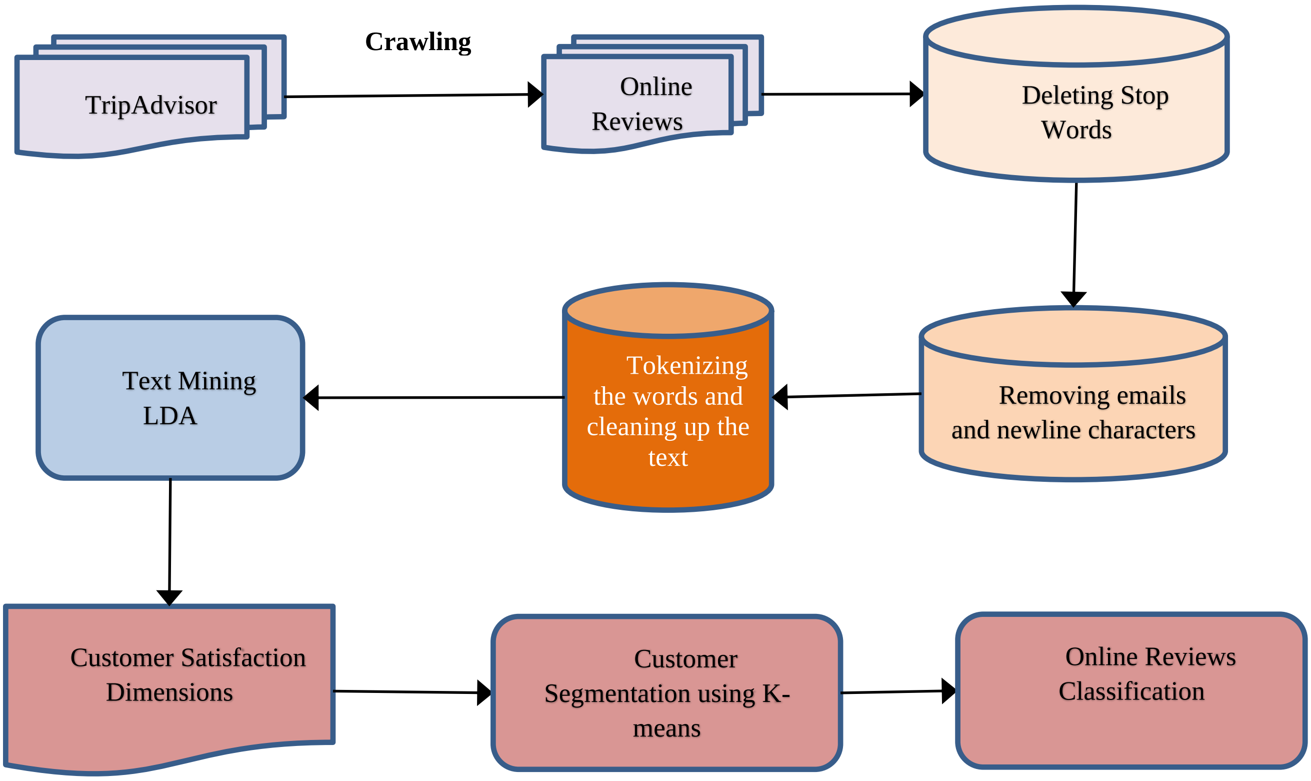

TripAdvisor
Crawling
Online Reviews
Deleting Stop Words
Removing emails and newline characters
Tokenizing the words and cleaning up the text
Text Mining LDA
Customer Satisfaction Dimensions
Customer Segmentation using K-means
Online Reviews Classification

Figure 7. Research Method

4. Data Analysis

Online reviews data collected from TripAdvisor were preprocessed and meaningless words were excluded. LDA is utilized to discover the satisfaction dimensions of customers' online reviews. Gensim library was utilized to create LDA topic modeling method. The analysis of the data entails four main stages; the cleaning and preparation of the social data; the discovery of satisfaction dimensions from the online reviews, customer segmentation based on criteria ratings, and the visualizations of the dimensions. A stage of online reviews classification is provided in this study as well. Online reviews are usually short and the LDA is limited in handling short textual data (Zhang et al., 2021). Hence, a preprocessing stage of the data is essential to improve the performance of the generative model. The preprocessing of the data includes (1) removing the stop words, (2) removing emails and newline characters, (3) tokenizing the words, and (4) cleaning up the text. The stop word list provided by the NLTK package is extended by adding more stop words to the list. The Python package; pyLDAvis was deployed to present the visualizations of the LDA topics. The number of topics was adjusted until we obtained non-overlapping segments of data, which leads to 4 main topics as presented in Figure 3. Besides, to visualize the topics we generated a word cloud of each topic as presented in Figure 9. The circles in Figure 8 represent the topics, in which the size of the topic demonstrates its significance. Figure 9, presents the 4 main topics in the online reviews dataset and the most relevant word distribution related to a specific topic. In this study, 4 topics are generated and 30 keynote words for each topic are obtained. The data cloud for each topic is presented in Figure 9, presenting the higher 15 words in terms of frequency in that specific topic.

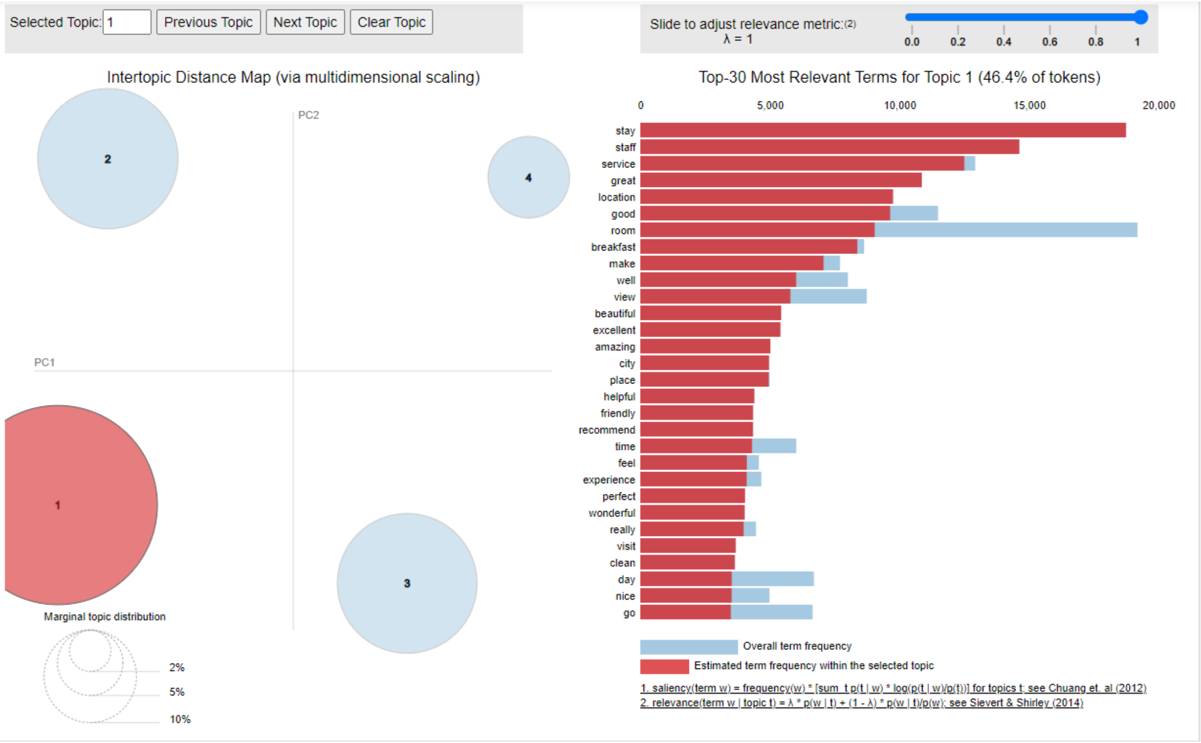

Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (46.4% of tokens)
0
5,000
10,000
15,000
20,000
stay
staff
service
great
location
good
room
breakfast
make
well
view
beautiful
excellent
amazing
city
place
helpful
friendly
recommend
time
feel
experience
perfect
wonderful
really
visit
clean
day
nice
go
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8. The Main Extracted Topics

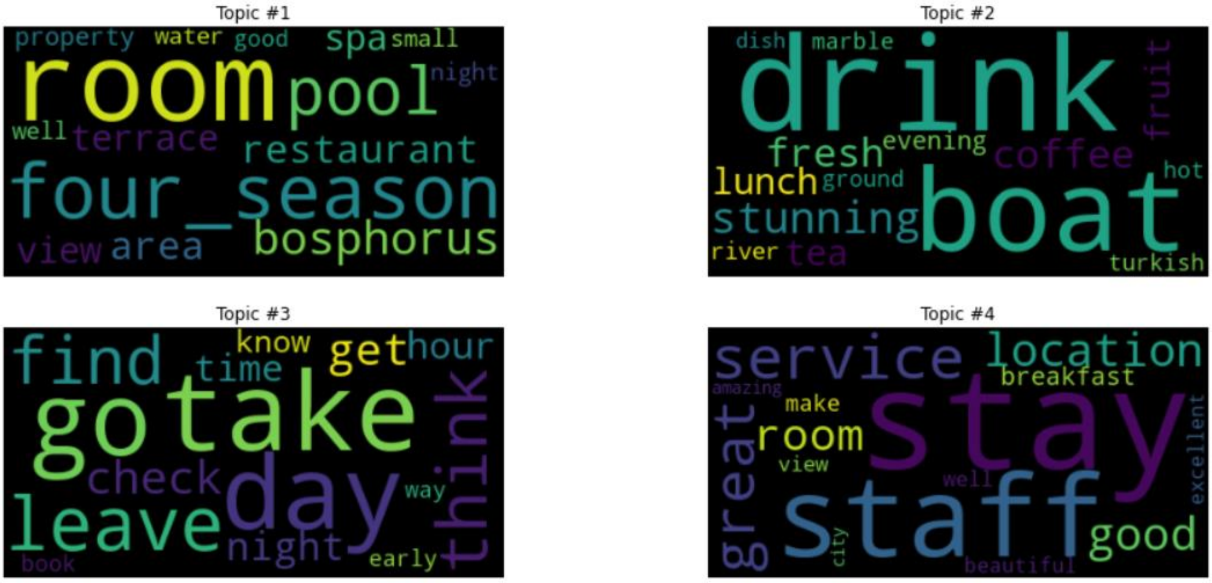


Figure 9. Generated Word Cloud

K-means clustering was deployed to separate the customers into groups with similar ratings. In the clustering approach, 3 segments (k = 3) were regarded. The outcomes of k-means clustering are illustrated in Table 2. The Segment1, Segment2, and Segment3 centroids are [27.392175, 36.879851, 23.958665, 33.388009, 37.600293, 35.423475], [40.050183, 42.673045, 43.954379, 42.569097, 46.634343, 42.353479], and [47.844337, 48.088739, 49.676296, 47.976478, 48.700235, 47.985772] respectively. Dividing the customers into segments with similar tendencies through their ratings is important to gain deep insight into the customers' preferences. Along with that, new customers can be assigned to a segment based on the distance of their ratings to the clusters' centroid. The obtained centroid centers reveal the dimensions which have more impact on customers' satisfaction. For example, the obtained results in Table 2 show that in Segment 1, the customers' ratings in cleanliness criteria are higher compared with others in the same group. In Segment 2, it is obvious that the customers provided moderate criteria ratings for the entire group, the customers have given lower ratings for value criteria than other criteria, therefore, they have indicated their less satisfaction related to value criteria. It is clear that the travelers' satisfaction with cleanliness criteria is high compared with other criteria ratings in Segment 2. In Segment 3, customers' ratings are high in general throughout the group. In segment 3, It is clear that the travelers have been notably happy with the service of the obtained data from the targeted green hotels.

Table 2. Cluster centroids

Attribute	Segment1	Segment2	Segment3
Value	27.392175	40.050183	47.844337
Location	36.879851	42.673045	48.088739
Service	23.958665	43.954379	**49.676296**
Rooms	33.388009	42.569097	47.976478
Cleanliness	**37.600293**	**46.634343**	48.700235
Sleep Quality	35.423475	42.353479	47.985772

LSTM and GRU were deployed to classify customers' reviews into either positive or negative. In both LSTM and GRU, adam optimizer, and sigmoid activation function were used. The used number of epochs for training the model is 15 epochs. The training and validation loss curve of the LSTM model is presented in Figure 10. The accuracy curve of the LSTM model is illustrated in Figure 11. The accuracy obtained for the LSTM model was 0.8670 and the obtained loss was 0.3297. The precision, recall, and f-1 score of the LSTM model was provided in Table 3. It is obvious that the curve of training was decreasing towards zero in the training and validation loss, and increasing towards 1 in the accuracy in both LSTM and GRU models.

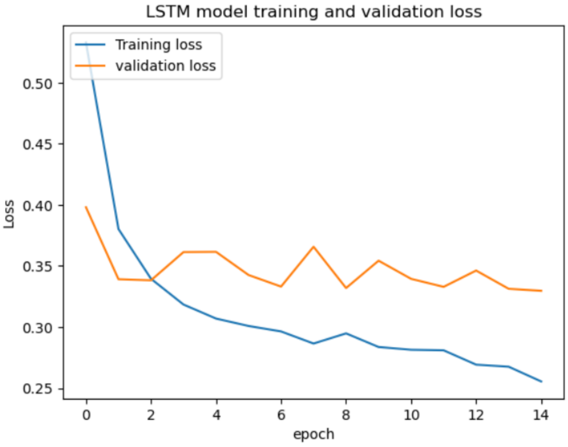


Figure 10. LSTM model loss curve

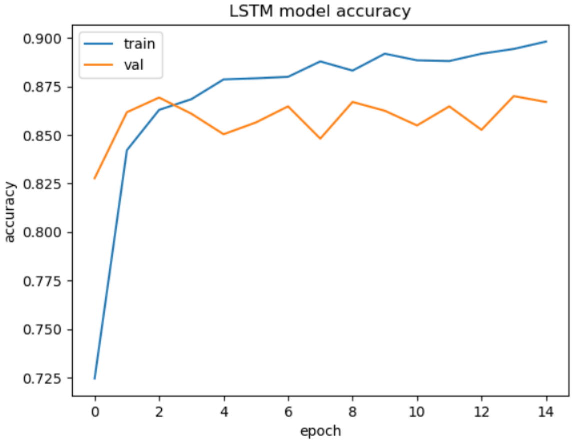


Figure 11. LSTM accuracy curve

Table 3. Results of the LSTM model

Category	Precision	Recall	F-1 score
Negative	0.87	0.87	0.87
Positive	0.86	0.86	0.86

GRU model was trained using the collected balanced dataset as well. GRU model training and validation loss are depicted in Figure 12. GRU model accuracy is illustrated in Figure 13. The obtained accuracy in the case of GRU was 0.8700 and the obtained loss value was 0.3408. The precision, recall, and f-1 score of the trained GRU model was provided in Table 4.

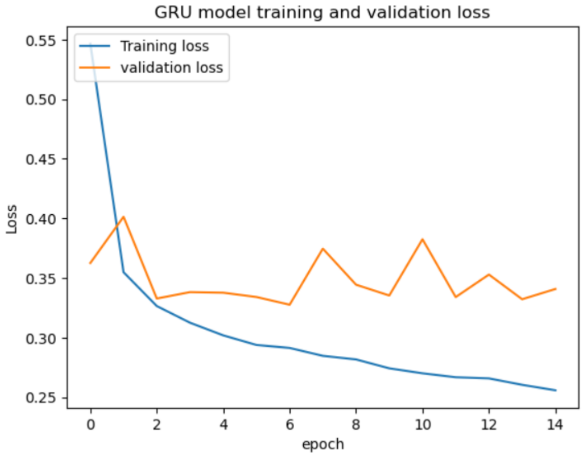


Figure 12. GRU model loss curve

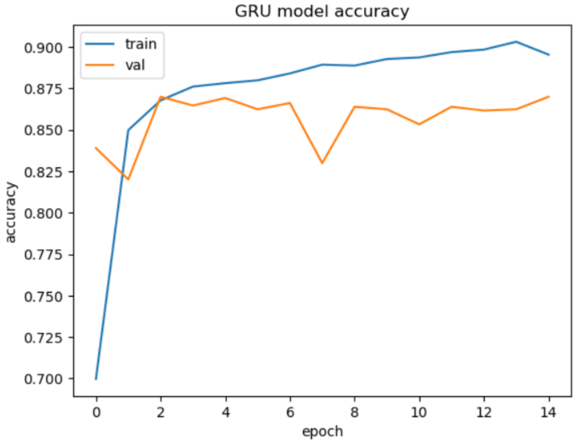


Figure 13. GRU model accuracy curve

Table 4. Results of the LSTM model

Category	Precision	Recall	F-1 score
Negative	0.88	0.87	0.87
Positive	0.86	0.87	0.87

.

5. Results and Discussion

Data collection and machine learning application on the dataset results reveal the main topics and the main features that impact the customers' attention. We modified the number of topics given to LDA until we found that 4 topics is the appropriate number of topics that showed non-intersecting clusters with adequate space between the topics reached. The results of the 4 main topics are depicted in Figure 8, and Figure 9. We can infer from Figure 9. the main criteria and main features that caught the customers' attention. Topic 1 focuses on properties like pools, restaurants, terrace views, etc., Topic 2 concentrates on beverage services like drinks, tea, coffee, etc., Topic 3 words cloud is centered on other quality features such as the time, hour, night, day, etc., and finally, Topic 4 considered general services like staff, breakfast, stay, etc. Following revealing satisfaction dimensions from the obtained dataset, customers were partitioned into 3 segments using the k-means clustering technique. The extracted segment revealed customers' behaviors for each rating criterion. We can infer from cluster centroids in Table 2. that customer satisfaction in Segment1, Segment2, and Segment3 are low, moderate, and high respectively. In Segment1, the customers' satisfaction has been high with cleanliness and location criteria, and low with the service criteria, Generally customers' satisfactiın in Segment1 is relatively low compared with the other two segments. In Segment2 the customers have been more satisfied with the cleanliness criteria and less satisfied with value criteria compared with the other criteria in the same group, and the customers' satisfaction in Segment2 is relatively moderate. Finally, in Segment3 the customers' satisfaction is high, especially for service criteria where the customers have been highly satisfied with the 49.676296 centroid center. cluster centroid. These presented features in the extracted 4 topics and the extracted centroids for each segment can help hotel managers and decision-makers to understand the customers' concerns about green hotels. Hotel managers and decision-makers can know by these topics and features what sections they should enhance in their eco-friendly hotels as well.

As presented in Figure 9, four main topics were extracted from the online reviews, we refer to Topic 1 as facilities-centered, Topic 2 as beverage-centered, Topic 3 as timing-centered, and finally, Topic 4 as services-centered. From this clustering, we can confirm the alignment of our findings with the results of previous literature in similar contexts. The facilities in the hotels such as rooms, pools, and restaurants are vital for the choice of the hotel and the assessment of the overall tourism experience (Bauer et al., 1993). Beverages and drinks also gained the interest of researchers in the tourism and hospitality businesses (Park et al., 2016; Türker & Süzer, 2022). Timing in terms of services such as check out, dining, and room services is important for tourist satisfaction and has been explored in previous literature (De Palma et al., 2018). Finally, the important role of service quality in the hotel industry has been endorsed in previous literature through empirical outcomes (Fan et al., 2022; Harif et al., 2022; Perramon et al., 2022). As presented in table 2. customer segmentation has been utilized in several researches due to the significance of centroids in the prediction of new customers' satisfaction by means of their criteria ratings (Nilashi, Abumalloh, Alghamdi, et al., 2021; Nilashi et al., 2022).

In this study, in order to classify the customers' reviews in terms of being positive reviews or negative we trained LSTM and GRU models. From the results, we can infer that in this experiment GRU model gives an accuracy of 0.8700 which was higher than the accuracy of LSTM with 0.8670. The model shows successful results which is able to recognize whether a customer's review is negative or positive with a high rate of accuracy. LSTM, and GRU which is an extension of RNN that solved the problem of gradient vanishing play a key role and had been widely utilized in text classification in the literature (Liang & Niu, 2022; Moirangthem & Lee, 2021; Wadud et al., 2022).

Our results support the findings of previous literature in the context of green tourism as these factors were located as important drivers of customers satisfaction in the study by (Bauer et al., 1993; D'Alessandro, 2016; De Palma et al., 2018; Kim et al., 2016; Park et al., 2016; Zamparini et al., 2022).

6. Conclusion

Revealing customers' expectations are pretty important for the tourism sector and particularly for green tourism practices in hotels. In this study, we collected online review data which is considered an important type of big social data generated by users on the TripAdvisor site using a crawling technique that crawled online reviews from hotel sites using their URLs. Gathered data preprocessed, stop words were deleted and extended, extended stop words contained meaningless words and repeated words, emails and newlines were removed, words were tokenized, and the text was cleaned. The most important features that gained tourists' interest were discovered by utilizing the LDA topic modeling technique. In order to understand the customers' behaviors better we partitioned the customers into 3 main groups using the k-means clustering technique. Finally, a new balanced dataset was built in order to be utilized in the classification model. After the data preprocessing stage LSTM and GRU were trained, and GRU was given higher accuracy. Consequently, GRU was deployed.

Traveler satisfaction is fundamentally significant in the tourism sector and particularly in environment-friendly hotels. This study utilized online reviews in echo-friendly hotel sites and applied natural language processing techniques on the online reviews to discover the travelers' satisfaction dimensions. Research findings show t 4 major satisfaction dimensions that we covered in the discussion section. These dimensions are highly important for green hotels to take into consideration. Hotels can enhance the main features in their area based on these dimensions extracted from travelers' online reviews. The research presented insights for decision-makers in the tourism industry by revealing the important factors for tourists' experiences and clustering the customers with similar rating behavior.

7. Limitation of Study and Future Work

The study has a few limitations in terms of the collected data and the deployed method. The data was collected from one online social platform; TripAdvisor; regarding its popularity among tourists, other portals can be utilized to investigate tourists' perceptions more broadly. The deployed method focused on discovering the dimensions of travelers' satisfaction using the LDA, segmenting the customers into groups with similar rating behavior, and training comparing the proposed ML models for the classification of the collected customers' reviews. other research models that entail a survey-based approach, and customer prediction using fuzzy logic approach can present a wider perception of the ranking of the importance levels of the discovered satisfaction dimensions

References

Acampora, A., Lucchetti, M. C., Merli, R., & Ali, F. 2022. The theoretical development and research methodology in green hotels research: A systematic literature review. Journal of Hospitality and Tourism Management, 51, 512-528.

Afrizal, A. D., Rakhmawati, N. A., & Tjahyanto, A. 2019. New filtering scheme based on term weighting to improve object based opinion mining on tourism product reviews. Procedia Computer Science, 161, 805-812.

Alzate, M., Arce-Urriza, M., & Cebollada, J. 2022. Mining the text of online consumer reviews to analyze brand image and brand positioning. Journal of Retailing and Consumer Services, 67, 102989.

Arulraj, T., & Daisy, S. J. S. 2021. Mining online review for predicting sales performance. Materials Today: Proceedings, 47, 93-99.

Association, G. H. 2008. What are green hotels. Retrieved May, 10, 2008.

Bauer, T., Jago, L., & Wise, B. 1993. The changing demand for hotel facilities in the Asia Pacific region. International Journal of Hospitality Management, 12(4), 313-322.

Berezan, O., Raab, C., Yoo, M., & Love, C. 2013. Sustainable hotel practices and nationality: The impact on guest satisfaction and guest intention to return. International Journal of Hospitality Management, 34, 227-233.

Bian, Y., Ye, R., Zhang, J., & Yan, X. 2022. Customer preference identification from hotel online reviews: A neural network based fine-grained sentiment analysis. Computers & Industrial Engineering, 108648.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bozkurt, F., Çoban, Ö., Baturalp Günay, F., & Yücel Altay, Ş. 2019. High performance twitter sentiment analysis using CUDA based distance kernel on GPUs. Tehnički vjesnik, 26(5), 1218-1227.

Chen, Q., Hu, M., He, Y., Lin, I., & Mattila, A. S. 2022. Understanding guests' evaluation of green hotels: The interplay between willingness to sacrifice for the environment and intent vs. quality-based market signals. International Journal of Hospitality Management, 104, 103229.

Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. 2018. Leveraging social media news to predict stock index movement using RNN-boost. Data & Knowledge Engineering, 118, 14-24.

Chimmula, V. K. R., & Zhang, L. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135, 109864.

D'Alessandro, F. 2016. Green Building for a Green Tourism. A new model of eco-friendly agritourism. Agriculture and agricultural science procedia, 8, 201-210.

De Palma, A., Criado, C. O., & Randrianarisoa, L. M. 2018. When Hotelling meets Vickrey. Service timing and spatial asymmetry in the airline industry. Journal of Urban Economics, 105, 88-106.

Fan, H., Gao, W., & Han, B. 2022. How does (im) balanced acceptance of robots between customers and frontline employees affect hotels' service quality? Computers in Human Behavior, 133, 107287.

Filimonau, V., Matute, J., Mika, M., Kubal-Czerwińska, M., Krzesiwo, K., & Pawłowska-Legwand, A. 2022. Predictors of patronage intentions towards 'green'hotels in an emerging tourism market. International Journal of Hospitality Management, 103, 103221.

Godnov, U., & Redek, T. 2016. Application of text mining in tourism: case of Croatia. Annals of Tourism Research, 58, 162-166.

Han, H., Lee, J.-S., Trang, H. L. T., & Kim, W. 2018. Water conservation and waste reduction management for increasing guest loyalty and green hotel practices. International Journal of Hospitality Management, 75, 58-66.

Harif, M. A. A. M., Nawaz, M., & Hameed, W. U. 2022. The role of open innovation, hotel service quality and marketing strategy in hotel business performance. Heliyon, e10441.

Hochreiter, S., & Schmidhuber, J. 1996. LSTM can solve hard long time lag problems. Advances in neural information processing systems, 9.

Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9(8), 1735-1780.

Huang, J., Guo, Y., Wang, C., & Yan, L. 2019. You touched it and I'm relieved! The effect of online review's tactile cues on consumer's purchase intention. Journal of Contemporary Marketing Science.

Huang, S., Zhang, J., Yang, C., Gu, Q., Li, M., & Wang, W. 2022. The interval grey QFD method for new product development: Integrate with LDA topic model to analyze online reviews. Engineering Applications of Artificial Intelligence, 114, 105213.

Jung, M., Lee, H., & Tani, J. 2018. Adaptive detrending to accelerate convolutional gated recurrent unit training for contextual video recognition. Neural Networks, 105, 356-370.

Khaldi, R., El Afia, A., Chiheb, R., & Tabik, S. 2023. What is the best RNN-cell structure to forecast each time series behavior? Expert Systems with Applications, 215, 119140.

Kim, J.-Y., Hlee, S., & Joun, Y. 2016. Green practices of the hotel industry: Analysis through the windows of smart tourism system. International Journal of Information Management, 36(6), 1340-1349.

Liang, M., & Niu, T. 2022. Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs. Procedia Computer Science, 208, 460-470.

Ma, G., Ma, J., Li, H., Wang, Y., Wang, Z., & Zhang, B. 2022. Customer behavior in purchasing energy-saving products: Big data analytics from online reviews of e-commerce. Energy Policy, 165, 112960.

Magoulas, R., & Swoyer, S. 2020. AI Adoption in the Enterprise. Beijing: O′ Reilly. Recuperado de http://www. oreilly. com/data/free/ai ….

Merli, R., Preziosi, M., Acampora, A., & Ali, F. 2019. Why should hotels go green? Insights from guests experience in green hotels. International Journal of Hospitality Management, 81, 169-179.

Moirangthem, D. S., & Lee, M. 2021. Hierarchical and lateral multiple timescales gated recurrent units with pre-trained encoder for long text classification. Expert Systems with Applications, 165, 113898.

Nilashi, M., Abumalloh, R. A., Alghamdi, A., Minaei-Bidgoli, B., Alsulami, A. A., Thanoon, M., Asadi, S., & Samad, S. 2021. What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis. Telematics and Informatics, 64, 101693.

Nilashi, M., Abumalloh, R. A., Almulihi, A., Alrizq, M., Alghamdi, A., Ismail, M. Y., Bashar, A., Zogaan, W. A., & Asadi, S. 2021. Big social data analysis for impact of food quality on travelers' satisfaction in eco-friendly hotels. ICT Express.

Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Zogaan, W. A., Alhargan, A., Mohd, S., Azhar, S. N. F. S., Asadi, S., & Samad, S. 2022. Revealing travellers' satisfaction during COVID-19 outbreak: moderating role of service quality. Journal of Retailing and Consumer Services, 64, 102783.

Nilashi, M., Minaei-Bidgoli, B., Alrizq, M., Alghamdi, A., Alsulami, A. A., Samad, S., & Mohd, S. 2021. An analytical approach for big social data analysis for customer decision-making in eco-friendly hotels. Expert Systems with Applications, 186, 115722.

Park, S., Lundeen, E., & Blanck, H. 2016. Knowledge of Health Conditions Related to Drinking Sugar-Sweetened Beverage and Sugar-Sweetened Beverage Intake Among US Adults. Journal of Nutrition Education and Behavior, 48(7), S98.

Perramon, J., Oliveras-Villanueva, M., & Llach, J. 2022. Impact of service quality and environmental practices on hotel companies: An empirical approach. International Journal of Hospitality Management, 107, 103307.

Prihayati, Y., & Veriasa, T. O. 2021. Developing green tourism to create the sustainable landscape: evidence from Community-based Coffee Tourism (CbCT) in Puncak, Bogor, Indonesia. IOP Conference Series: Earth and Environmental Science,

Rita, P., Moro, S., & Cavalcanti, G. 2022. The impact of COVID-19 on tourism: Analysis of online reviews in the airlines sector. Journal of Air Transport Management, 104, 102277.

Shaheen, M., Zeba, F., Chatterjee, N., & Krishnankutty, R. 2019. Engaging customers through credible and useful reviews: the role of online trust. Young Consumers.

Sim, Y., Lee, S. K., & Sutherland, I. 2021. The impact of latent topic valence of online reviews on purchase intention for the accommodation industry. Tourism Management Perspectives, 40, 100903.

Streimikiene, D., Svagzdiene, B., Jasinskas, E., & Simanavicius, A. 2021. Sustainable tourism development and competitiveness: The systematic literature review. Sustainable development, 29(1), 259-271.

Tian, Y., & Zhang, Y. 2022. Pricing of crowdfunding products with strategic consumers and online reviews. Electronic Commerce Research and Applications, 54, 101169.

Tuna, H., & Başdal, M. 2021. Curriculum evaluation of tourism undergraduate programs in Turkey: A CIPP model-based framework. Journal of Hospitality, Leisure, Sport & Tourism Education, 29, 100324.

Türker, N., & Süzer, Ö. 2022. Tourists' food and beverage consumption trends in the context of culinary movements: The case of Safranbolu. International Journal of Gastronomy and Food Science, 27, 100463.

UNEP. 2013. World's Largest Travel Site Awards Qualifying Accommodations Across the U.S. with Bronze, Silver, Gold or Platinum Status. Retrieved October from https://www.unep.org/es/node/6002

Verma, V. K., Chandra, B., & Kumar, S. 2019. Values and ascribed responsibility to predict consumers' attitude and concern towards green hotel visit intention. Journal of Business Research, 96, 206-216.

Wadud, M. A. H., Kabir, M. M., Mridha, M., Ali, M. A., Hamid, M. A., & Monowar, M. M. 2022. How can we manage offensive text in social media-a text classification approach using LSTM-BOOST. International Journal of Information Management Data Insights, 2(2), 100095.

Wang, Q., Zhang, W., Li, J., Mai, F., & Ma, Z. 2022. Effect of online review sentiment on product sales: The moderating role of review credibility perception. Computers in Human Behavior, 133, 107272.

Wei, X., & Taecharungroj, V. 2022. How to improve learning experience in MOOCs an analysis of online reviews of business courses on Coursera. The International Journal of Management Education, 20(3), 100675.

Williams, T., & Betak, J. 2018. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. Procedia Computer Science, 130, 98-102.

Wu, H., Zhang, Z., Li, X., Shang, K., Han, Y., Geng, Z., & Pan, T. 2022. A novel pedal musculoskeletal response based on differential spatio-temporal LSTM for human activity recognition. Knowledge-Based Systems, 110187.

Wu, L., & Noels, L. 2022. Recurrent Neural Networks (RNNs) with dimensionality reduction and break down in computational mechanics; application to multi-scale localization step. Computer Methods in Applied Mechanics and Engineering, 390, 114476.

Xianghua, F., Guo, L., Yanyan, G., & Zhiqiang, W. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems, 37, 186-195.

Yeşiltaş, M., Gürlek, M., & Kenar, G. 2022. Organizational green culture and green employee behavior: Differences between green and non-green hotels. Journal of Cleaner Production, 343, 131051.

Yu, M., Cheng, M., Yang, L., & Yu, Z. 2022. Hotel guest satisfaction during COVID-19 outbreak: The moderating role of crisis response strategy. Tourism Management, 93, 104618.

Zamparini, L., Domènech, A., Miravet, D., & Gutiérrez, A. 2022. Green mobility at home, green mobility at tourism destinations: A cross-country study of transport modal choices of educated young adults. Journal of Transport Geography, 103, 103412.

Zhang, C., Peng, K., Dong, J., & Miao, L. 2022. A comprehensive operating performance assessment framework based on distributed Siamese gated recurrent unit for hot strip mill process. Applied Soft Computing, 109889.

Zhang, E., Li, H., Huang, Y., Hong, S., Zhao, L., & Ji, C. 2022. Practical multi-party private collaborative k-means clustering. Neurocomputing, 467, 256-265.

Zhang, N., Liu, R., Zhang, X.-Y., & Pang, Z.-L. 2021. The impact of consumer perceived value on repeat purchase intention based on online reviews: by the method of text mining. Data Science and Management, 3, 22-32.

Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., & Wang, J. 2017. Machine health monitoring using local feature-based gated recurrent unit networks. IEEE Transactions on Industrial Electronics, 65(2), 1539-1548.

Zibarzani, M., Abumalloh, R. A., Nilashi, M., Samad, S., Alghamdi, O., Nayer, F. K., Ismail, M. Y., Mohd, S., & Akib, N. A. M. 2022. Customer satisfaction with Restaurants Service Quality during COVID-19 outbreak: A two-stage methodology. Technology in Society, 70, 101977.