



Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca

Ahmed Alsayat¹

Received: 15 August 2021 / Accepted: 21 October 2022 / Published online: 28 October 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

Big social data and user-generated content have emerged as important sources of timely and rich knowledge to detect customers' behavioral patterns. Revealing customer satisfaction through the use of user-generated content has been a significant issue in business, especially in the tourism and hospitality context. There have been many studies on customer satisfaction that take quantitative survey approaches. However, revealing customer satisfaction using big social data in the form of eWOM (electronic word of mouth) can be an effective way to better understand customers' demands. In this study, we aim to develop a hybrid methodology based on supervised learning, text mining, and segmentation machine learning approaches to analyze big social data on travelers' decision-making regarding hotels in Mecca, Saudi Arabia. To do so, we use support vector regression with sequential minimal optimization (SMO), latent Dirichlet allocation (LDA), and *k*-means approaches to develop the hybrid method. We collect data from travelers' online reviews of Mecca hotels on TripAdvisor. The data are segmented, and travelers' satisfaction is revealed for each segment based on their online reviews of hotels. The results show that the method is effective for big social data analysis and traveler segmentation in Mecca hotels. The results are discussed, and several recommendations and strategies for hotel managers are provided to enhance their service quality and improve customer satisfaction.

Keywords Text mining · Machine learning · Big social data · eWOM · Customer satisfaction · Segmentation · Hotel industry · Customer decision-making

1 Introduction

Traveler satisfaction, personalized service and a better hotel experience are among the top priorities for hotel managers, especially since good hotel service quality enhances travelers' loyalty and influences their repurchase intentions [1, 2]. Most international travelers place a high value on the quality of their accommodations [3]. The choice of hotel also serves as an example of a difficult decision-making process [4]. Thus, strategic planning and decision-making are critical components of business success, and having a thorough understanding of travelers' preferences can assist hotel managers in these endeavors

[5]. According to previous research, personal attitudes toward service are affected by customer satisfaction. In turn, it has been revealed that there is a link between attitudes and travel choices, with factors such as hotels' performance criteria playing an important role [6].

In the age of social media, consumers of products and services prefer to read feedback from other consumers before making a purchasing decision [7]. Customer feedback is vital for business growth, customer experience improvement, and new product and service development. In the tourism and hospitality context, it is critical for hotel managers to satisfy travelers and translate this satisfaction into behavioral commitment to remain competitive [8]. It is possible to assess and respond to customer satisfaction levels as well as behavioral intentions using a variety of methods. Customer feedback is commonly used by managers to identify future managerial objectives and to monitor a company's performance through customer satisfaction scores such as average customer satisfaction

✉ Ahmed Alsayat
asayat@ju.edu.sa

¹ Department of Computer Science, College of Computer and Information Sciences, Jouf University, 72388 Sakaka, Kingdom of Saudi Arabia

scores. Online customer reviews and ratings are the most accessible and prevalent form of electronic word of mouth. The accessibility of travel websites facilitates customer feedback, and travelers are asked to provide both quantitative and qualitative reviews of hotels, destinations, and other travel experiences. According to previous research, over 30% of internet users rate products online [9]. According to previous research, 82% of US adults read online reviews and ratings before buying a new product or service [10].

Due to the exponential growth in the volume of data transmitted on social networking sites and the recognition of data as an economic asset, efforts are being made to actively utilize big social data in a variety of businesses (e.g., tourism and hospitality) and contexts (e.g., customer satisfaction and preference learning). Although many studies have been conducted on the customer satisfaction and preference learning, only a few researchers have investigated satisfaction from the viewpoint of travelers through big social data [7, 11, 12]. Unlike traditional questionnaires, user-generated content is considered free of sample bias, vast, and objective. Travelers can share knowledge, opinions, and information about services, products, and brands online. The online travel content created or posted by users is increasingly read and shared by travelers. These data are also easier for businesses to monitor and manage because they are easily accessible and low cost. Thus, more tourism researchers in recent years have relied on user-generated content to reveal customers' perceptions of service quality and their level of satisfaction [13]. These studies can be divided into two types based on user-generated content: studies based on textual reviews and studies based on numerical data. Text mining is mainly used to extract useful information (e.g., satisfaction dimensions) from textual reviews [9, 14–16]. However, studies based on numerical data mainly rely on the use of clustering and prediction learning approaches [17].

According to previous studies, travelers' preferences and satisfaction can be measured using user-generated content [18–20]. Additionally, massive user-generated content can be easily collected online, covering various travelers, regions, and hotel types. Thus, user-generated content is a promising data source for exploring the impacts of performance criteria (e.g., location and rooms) on travelers' satisfaction. In regard to explaining customer behavior and satisfaction through big social data, machine learning techniques have proven to be effective [21]. Similar to previous studies, this research aims to employ machine learning techniques to analyze online customer reviews of hotel services in Mecca, Saudi Arabia. Accordingly, a new method using clustering and supervised learning with the aid of the text mining approach is developed. The method is able to analyze both qualitative

(textual reviews) and quantitative (numerical ratings) data. The contributions of this work are as follows:

- i. This study investigates an important issue in the hospitality context: market segmentation through the use of online customer reviews. Market segmentation has been an effective and powerful marketing approach for decades in many business contexts [22–24], especially in tourism and hospitality. The concept of market segmentation has been widely accepted by the tourism industry [25–27]. Big social data analysis has been an important topic for customer satisfaction analysis and segmentation [28, 29]. Many studies on customer satisfaction through qualitative and quantitative survey approaches have been conducted. However, revealing customer satisfaction through the use of big social data in the form of eWOM (electronic word of mouth) can be an effective way to better understand customers' demands [30]; through eWOM analysis, customers' preferences can be better detected. Therefore, this study is the first in the tourism and hospitality context to examine customer satisfaction through an online review analysis of hotels in Mecca, Saudi Arabia. The results demonstrate that the proposed method can provide valuable results from online reviews for customer satisfaction.
- ii. Although there have been many studies regarding customer segmentation in the hospitality context, in the case of Mecca hotels on TripAdvisor, there have been no studies on customer satisfaction and preference learning. Previous studies rely on qualitative or quantitative analysis through the use of survey-based questionnaires. In addition, eWOM analysis has rarely been investigated by previous research for customer satisfaction with Mecca hotels on TripAdvisor. Accordingly, a new method for big social data analysis is developed using supervised learning, text mining, and segmentation machine learning approaches to analyze big social data on travelers' decision-making regarding hotels in Mecca, Saudi Arabia. Specifically, we use support vector regression (SVR), latent Dirichlet allocation (LDA), and k -means approaches to develop the hybrid method. To better predict customer satisfaction, SVR is optimized by sequential minimal optimization (SMO). The data are collected from travelers' online reviews of Mecca hotels on TripAdvisor. The data are segmented, and the travelers' satisfaction is revealed for each segment based on their online reviews of hotels.
- iii. Evaluation of the market segmentation method is an important task. The method proposed in this study is

evaluated through customers' reviews of Mecca hotels on TripAdvisor. We perform several experiments on and evaluations of real-world data and compare the results with those of other prediction techniques. This comparison is performed to show the effectiveness of the clustering approach for customer segmentation. Specifically, we compare the results of evaluation methods with a multilayer perceptron artificial neural networks (MLP-ANNs), multiple linear regression (MLR), and decision trees. In this comparison, the MLP is optimized with the whale optimization algorithm (WOA).

The remainder of this work is structured as follows. In Sect. 2, we present previous works on online review analysis, customer segmentation and tourism in Saudi Arabia. In Sect. 3, the proposed hybrid method of analysis is presented. In Sect. 4, data collection and analysis are presented. In Sect. 5, discussions of the results and research implications are presented. Finally, this paper is concluded in Sect. 6. The abbreviations used in this research are provided in Table 1.

Table 1 Abbreviations used in this research

Acronyms	Description
eWOM	Electronic word of mouth
GDP	Gross domestic product
GPHC	Gaussian peak heuristic-based clustering
K-NN	K-nearest neighbors
LDA	Latent Dirichlet allocation
MAE	Mean absolute error
MCDM	Multi-criteria decision-making
MLP-ANN	Multilayer perceptron artificial neural network
MLR	Multiple linear regression
NN	Neural network
PCA	Principal component analysis
RBF	Radial basis function
RF	Random forest
RFM	Recency, frequency, and monetary
RMSE	Root mean square error
SMO	Sequential minimal optimization
SNS	Social network site
SOM	Self-organizing map
STM	Structural topic model
SVM	Support vector machine
SVR	Support vector regression
WOA	Whale optimization algorithm

2 Literature review

2.1 Related work on online review analysis

There have been many studies on the analysis of online reviews in different contexts. In this section, we provide a discussion of the results of previous research (see Table 2).

Chang et al. [3] focused on social media analytics and provided a new framework for extracting and visualizing Hilton hotel reviews and ratings. They collected data from TripAdvisor. They used the support vector machine (SVM) technique for sentiment classification. Ahani et al. [17] focused on market segmentation through the use of spa hotel reviews and ratings. They developed a new method using higher order value decomposition, a self-organizing map (SOM), and decision trees. They used spa hotel reviews and ratings on TripAdvisor for data collection. They revealed that there is a significant relationship between hotels' performance criteria and travelers' satisfaction. Khorsand et al. [31] focused on the analysis of 8 machine learning models (naïve Bayes, a decision tree, logistic regression, K-nearest neighbors (KNN), a support vector machine, a random forest (RF), a NN, and gradient boosting) to predict rates of new users. They collected data on Tehran hotels from TripAdvisor. Nilashi et al. [32] focused on travelers' decision-making using preference learning for spa hotels in Malaysia. They used TripAdvisor for data collection. They developed a new method using clustering, dimensionality reduction, and decision trees. They performed the traveler segmentation using the *k*-means technique. Nilashi et al. [33] focused on travelers' decision-making using travelers' reviews on SNSs (social network sites). They collected data on hotels from TripAdvisor. They used decision trees and fuzzy logic approaches for method development. Arenas-Márquez et al. [34] focused on online review analysis to identify travel group-type topics. They collected data from TripAdvisor. Their method was developed through convolutional neural encoding. Fernandes et al. [35] focused on measuring restaurant performance by combining historical sales data with online reviews. They developed a new data-driven approach using variable correlation analysis as well as text mining and sentiment analysis. Taecharunroj et al. [36] focused on traveler review analysis for tourist attractions in Phuket, Thailand. They collected data from TripAdvisor. They developed a new method using LDA and naïve Bayes modeling. Gebbels, et al. [37] focused on evaluating customer restaurant reviews of Clink restaurants. They collected data from TripAdvisor and used a qualitative case-study methodology for data analysis. Bigne et al. [38] focused on evaluating the helpfulness of online reviews. They collected data from

Table 2 Previous studies on online review analysis

References	Context	Method	Dataset
[3]	Extracting and visualizing Hilton hotel reviews and ratings	Support vector machine for sentiment classification	TripAdvisor
[7]	Customer satisfaction analysis for Canary Islands hotels	Clustering and multi-criteria decision-making	TripAdvisor
[17]	Market segmentation for spa hotels reviews and ratings	Clustering, dimensionality reduction, and decision trees	TripAdvisor
[31]	Analysis of machine learning models to predict rates of new users for Tehran hotels	Naïve Bayes, a decision tree, logistic regression, KNN, a support vector machine, a random forest, a neural network, and gradient boosting	TripAdvisor
[32]	Travelers' decision-making using preference learning for Malaysian spa hotels	Clustering, dimensionality reduction, and decision trees	TripAdvisor
[33]	Travelers' decision-making using travelers' online reviews	Clustering, decision trees, and fuzzy logic	TripAdvisor
[34]	Online reviews analysis to identify travel group-type topics	Convolutional neural encoding	TripAdvisor
[35]	Measuring restaurant performance by combining historical sales data with online reviews TripAdvisor	Correlation analysis among variables, text mining and sentiment analysis	
[36]	Traveler review analysis for tourist attractions in Phuket, Thailand	LDA and naïve Bayes modelling	TripAdvisor
[37]	Evaluating customer restaurant reviews of Clink restaurants	Qualitative case-study methodology	TripAdvisor
[38]	Evaluating the helpfulness of online reviews	Deep learning	TripAdvisor
[39]	Analysis of online reviews to investigate the sense of place of the Las Vegas Strip	LDA	TripAdvisor
[40]	Differences between Booking.com and TripAdvisor in branding co-creation	Qualitative and quantitative analysis	TripAdvisor and Booking.com
[41]	Prioritizing improvements in hotel services	Topic modeling using LDA	TripAdvisor
[42]	Measuring service quality through the use of unstructured data	Topic modeling using latent structural topic models (STMs)	TripAdvisor

TripAdvisor. They adopted deep learning for sentiment analysis. Song et al. [39] used online reviews to investigate the sense of place of the Las Vegas Strip. They collected data from TripAdvisor. They developed a machine learning approach using LDA for topic modeling. Borges-Tiago et al. [40] focused on the differences between Booking.com and TripAdvisor in branding co-creation. The authors collected data from both Booking.com and TripAdvisor. They used qualitative and quantitative approaches for data analysis. Ahani et al. [7] conducted a study on customer satisfaction through the use of machine learning approaches, clustering and multi-criteria decision-making (MCDM). They collected data on Canary Island hotels from TripAdvisor. Zhang et al. [41] developed an online review-driven method to prioritize improvements in hotel services. They performed topic modeling using LDA on

TripAdvisor data. Korfiatis et al. [42] focused on measuring service quality through the use of unstructured data. The authors developed a topic modeling application by using STMs on online reviews by airline passengers on TripAdvisor.

2.2 Methods of customer segmentation

To achieve business objectives, a company can use customer segmentation to target marketing efforts and resources toward valuable and loyal customers [43]. Customer segmentation is the process of grouping all customers of a context into diverse and distinct groups based on shared characteristics such as patterns, interests, location, or demographics. Customer segments can be created using psychological behavioral, demographic, and

geographic data. There have been many studies on customer segmentation. To recognize profitable customers for more targeted communications, [44] looked at a few demographic variables (e.g., income, experience, education, gender, age) for customer segmentation. Rahim et al. [45] applied data modeling techniques and the RFM (recency, frequency, and monetary) model to detect customers' behavioral patterns. Their scheme was for customer classification through repurchase behavior recognition. Huseynov et al. [46] used several behavioral characteristics for customer segmentation. They used price payment, coupon redemption, product diversity, the shopping rate, and the refund rate to find effective segments from the available data. They found five main segments through the use of these behavioral characteristics for businesses to develop marketing strategies. Wang et al. [47] focused on psychological factors (e.g., attitudes, subjective norms, and personal norms) to identify customer segments through purchase intention. Yadegaridehkordi et al. [48] developed a new method of traveler segmentation through online customer reviews. They used MCDM and a clustering approach to segment travelers for eco-friendly hotels. Ahani et al. [17] focused on traveler segmentation for spa hotels. They performed traveler segmentation using machine learning approaches. Alkhayrat et al. [49] focused on telecom customer segmentation. The authors used deep learning and noise removal techniques to develop the segmentation approach. They used *k*-means for data clustering and principal component analysis (PCA) to reduce high-dimensional data. Sivaguru et al. [50] focused on dynamic customer segmentation through the use of modified dynamic fuzzy *c*-means clustering. Sun et al. [51] developed a heuristic clustering method for customer segmentation. They used Gaussian peak heuristic-based clustering (GPHC) and a standardized Gaussian distribution to perform numerical experiments. Wu et al. [52] focused on customer segmentation based on purchase behaviors and developed a method using an RFM model and the *k*-means algorithm. Wu et al. [53] developed a customer segmentation model using dynamic interval type-2 fuzzy logic and investigated its application in e-commerce. Akar et al. [54] focused on customer purchase intentions and segmentation during the COVID-19 outbreak. Vohra et al. [55] used *k*-means and SOM clustering based on the RFM model for customer segmentation.

2.3 Tourism in Saudi Arabia

Tourism is the most important part of any economy's service sector, accounting for the majority of revenue. Despite Saudi Arabia's oil wealth, international demand for petrochemical products is declining year by year, affecting the Saudi economy [56]. As a result, the economy must

diversify away from the oil sector to remain competitive. Saudi Arabia's future growth depends heavily on the development of its tourism industry [56]. It is one of the key pillars of Vision 2030's plan to diversify the economy and reduce reliance on oil. Saudi Arabia's tourism revenue reached 5 billion USD in December 2020, compared with 28 billion USD in 2019 (see Fig. 1). Saudi Arabia is well known throughout the world for the holy cities of Mecca and Madinah. Hajjis and visitors from all over the world travel to Mecca and Madinah every year to perform the Hajj and Umrah. The Hajj and Umrah pilgrimages are extremely important to Saudi Arabia's inbound tourism industry. The tourism revenue of the Kingdom of Saudi Arabia is increasing in absolute terms, but its contribution to the country's GDP is decreasing proportionally [56].

Saudi Arabia's economic development has benefited significantly from religious tourism [57]. During this time, worshipers are charged a fee. As foreigners, they will have to pay for food and housing, and accordingly, they can contribute to the country's foreign reserves and its income, which improves Saudis' quality of life. The government's Vision 2030 plan includes achieving sustainable economic growth [57]. According to studies, the number of people visiting Mecca has increased, and as a result, the city is playing a key role in Saudi Arabia's implementation of its Vision 2030 plan. Housing construction and other related infrastructure are also considered part of economic development. There has been a considerable amount of pressure to construct residential buildings that can house people during worship hours. There has been an increase in the growth and development of the hospitality industry as a result of the increase in the number of people traveling to Mecca and Madinah. As a result of international operations looking for opportunities and increasing the number of pilgrims in line with the Vision 2030, religious tourism is on the rise. According to industry experts, a significant increase in religious tourism will create enormous opportunities in Mecca's hotel and accommodation sectors. Furthermore, the transportation sector has grown to meet the needs of Hajj pilgrims and their belongings. Millions of Muslims from around the world travel to Saudi Arabia and Mecca each year to participate in the Holy Hajj, which takes place during the month of Dhu al-Hijjah. As a result, religious tourism remains a significant source of revenue for Saudi Arabia.

In 2017, approximately 19 million international tourists visited Saudi Arabia. According to the latest available statistics from the United Nations World Tourism Organization (UNWTO), the number of hotels and similar establishments reached approximately 8.3 thousand units in 2019, an increase from 5.5 thousand units in 2008. According to data published by the UNWTO, the total number of rooms in establishments in 2019 was

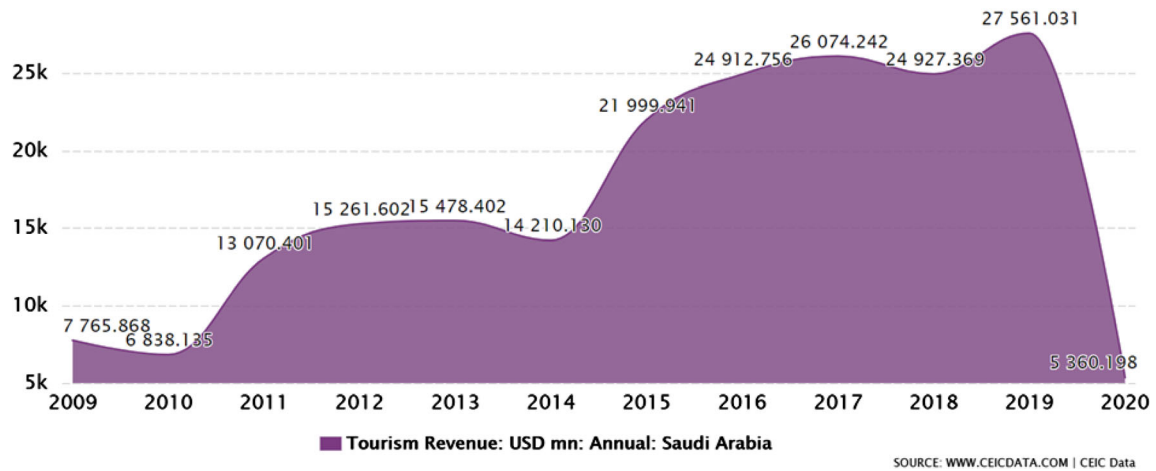


Fig. 1 Saudi Arabia's tourism revenue from 2004 to 2020

approximately 606.5 thousand rooms, up from 201.6 thousand rooms in 2008. Statistics also show that Mecca's contribution to Saudi Arabia's gross domestic product (GDP) from 2006 to 2016 was significant. According to the Saudi Arabian government, Mecca's direct contribution to Saudi Arabia's GDP in 2026 will be approximately 6.6 billion US dollars.

3 Method development

The aim of this study is to develop a method to segment travelers in hotels in Mecca, Saudi Arabia. Specifically, we aim to provide a hybrid method to reveal travelers' satisfaction through a set of hotel performance criteria. The method is developed using unsupervised clustering and supervised learning approaches. The proposed method is presented in Fig. 2. The steps of the proposed method are presented as follows:

Step 1: This research relied on online customer reviews. Accordingly, it was important to obtain data from a trusted website. The data were collected from TripAdvisor. This platform is widely used for data collection [40, 58, 59]. TripAdvisor provides complete forms for travelers to provide their experience with regard to the quality of services in restaurants and hotels [60]. In Step 1, data preprocessing was performed on the collected data. In this step, we tried to perform a data cleaning process on the textual and numerical ratings. In this stage, records with no criteria ratings were removed from the dataset. In addition, short reviews were not considered in the next stage of data analysis, and accordingly, we removed them from the dataset.

Step 2: In this step, we performed data segmentation through the use of unsupervised machine learning. It has

been shown that the clustering of data has a significant impact on the accuracy of supervised learning. This research relied on the *k*-means clustering technique [61–63]. The main advantage of *k*-means is its simplicity [64]. The computational simplicity of the *k*-means method makes it suitable for processing large datasets [65]. *K*-means always finds a local optimum for any given initial centroid locations [66]. Segmentation was performed for different *k* values, and the quality of clusters was evaluated through the use of the silhouette coefficient (SC) or silhouette score [67, 68].

Step 3: To discover the main satisfaction factors from the online reviews, in this step, text mining was performed on the collected data. This research relied on the LDA approach [69–71] to analyze textual data and perform topic modeling. LDA has been an effective approach for qualitative data analysis [72]. LDA is insensitive to the noise in the data and has powerful computing capabilities for big data [70]. In contrast to the Bayes classifier, which requires the classes to be predefined, LDA generates possible patterns from the data itself and is a data-driven technique. We performed text mining on each segment to reveal the dimensions of traveler satisfaction.

Step 4: This stage was performed after data segmentation. In this step, we performed supervised learning to predict travelers' satisfaction and identify their preferences for each segment. In this research, we relied on the SVR approach [73, 74]. We performed SVR to predict customer satisfaction from travelers' ratings of the hotels through a set of performance quality factors. In addition, we optimized SVR using the SMO solver algorithm [75, 76]. In this stage, the relationship between the performance quality factors and traveler satisfaction was examined.

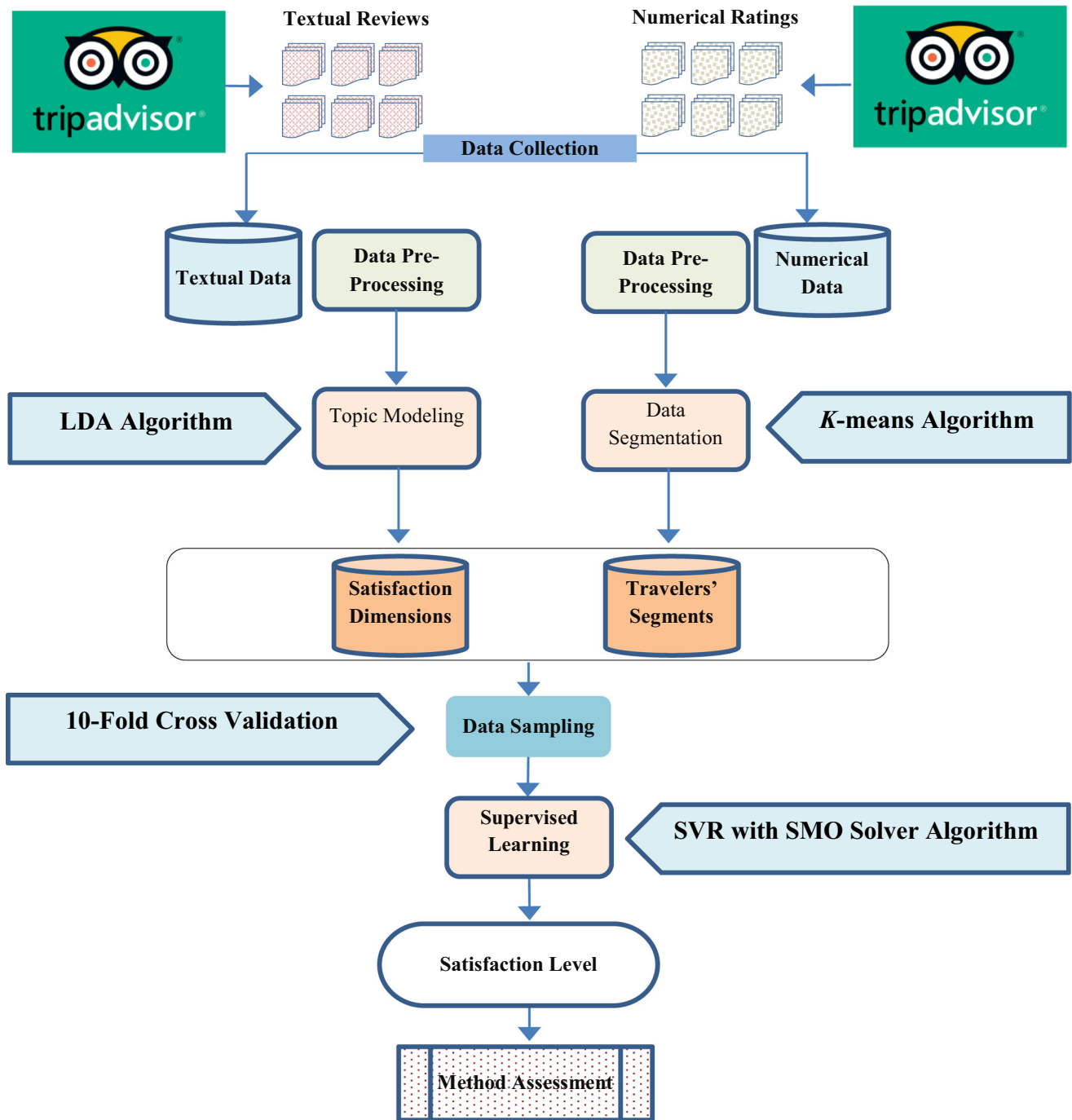


Fig. 2 Research method

Step 5: In this step, the prediction models generated by the SVR approach were evaluated. This research relied on several evaluation metrics, such as a sampling approach, the root mean square error (RMSE) [77–81], the mean absolute error (MAE) [77], and the coefficient of determination [82–84]. In addition, all models were evaluated by employing the tenfold cross-validation approach [85, 86].

3.1 K-means clustering technique

The *k*-means algorithm [87], which minimizes the sum of squared errors (SSE), is the most widely used clustering algorithm [88–90]. *K*-means is an iterative algorithm [91, 92] that can be described in the following steps, as shown in Fig. 3. The silhouette index (SI) is a reliable indicator for measuring clustering performance because it

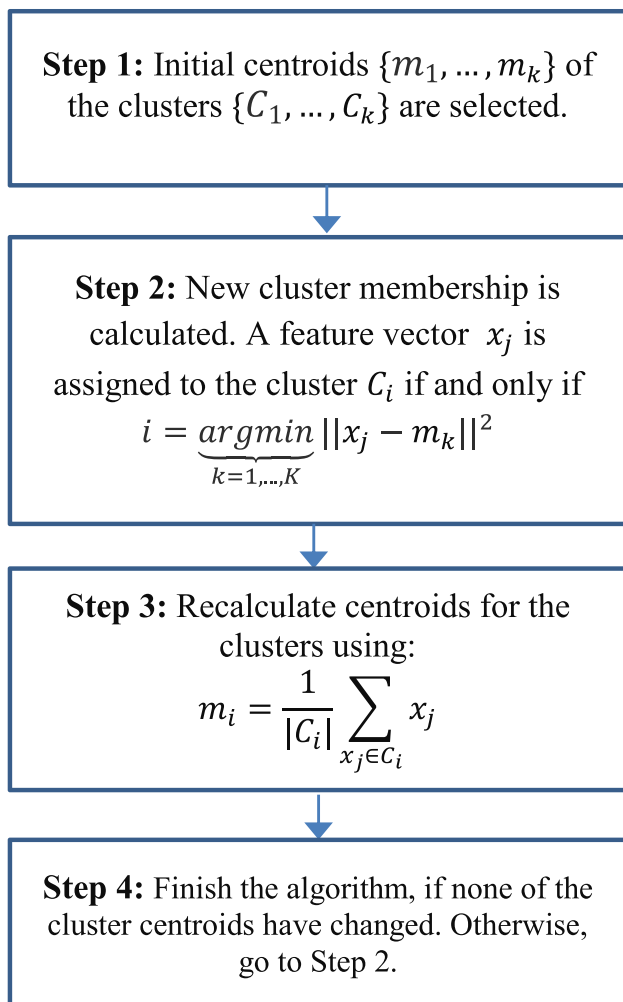


Fig. 3 K-means procedure

quantifies how similar an instance of a dataset is to instances within its own cluster versus instances in other clusters. Each instance j has a silhouette index SI_j . The silhouette index is calculated in a range of $[-1, +1]$. SI_j values close to -1 denote that an instance was probably assigned to the wrong cluster. SI_j values close to 0 denote that the instance does not clearly belong to one cluster or another. A value closer to $+1$ denotes that an instance was correctly assigned to its current cluster. SI_j is estimated in the following manner:

$$SI_j = \frac{b(j) - a(j)}{\max\{b(j), a(j)\}} \tag{1}$$

where $b(j)$ indicates the average distance from the j -th instance to all others assigned to the nearest neighboring segment and $a(j)$ represents the average distance from the j -th instance to all others in its segment. When calculating the distance between objects, Euclidean or Manhattan distances are generally utilized, and a silhouette graph can be constructed to demonstrate the performance of a

clustering technique. Averaging SI s over all n clustered observations can be performed to determine the global quality of a clustering technique. SI s are technique independent and can be used to find the optimal k , i.e., the number of clusters. The procedure for calculating the silhouette coefficient can be found at the following URL: <https://www.mathworks.com/help/stats/silhouette.html>.

3.2 Topic Modeling Using LDA

In this research, our aim from textual reviews is to examine the relationships between terms and extract the topics from their structures by using text mining with LDA topic modeling [70, 93–96]. In LDA, which is unsupervised and probabilistic in nature, it is assumed that all documents in a corpus are formed by a specific and manually definable number of topics. In LDA, each document is equally important and includes a “bag of words”. The words in each document are assumed to be unordered. A probability mass function of words is also used to describe a topic. To select topics, each document uses a probability mass function. The graphical representation of LDA is presented in Fig. 4. As illustrated, the schema contains three distinct levels that correspond to the corpus, the documents, and the terms.

In latent Dirichlet allocation, documents are seen as a distribution over the latent topics, which is described by Dirichlet distribution $\text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_k)$. A topic is selected based on the document-topics distribution, i.e., Multinomial(θ), where θ indicates the probability that a given document d belongs to a certain topic k . A Dirichlet distribution based on β encodes that a topic has a limited number of words. Given α and β , the topic mixture’s joint multivariate distribution, the set of N topics z , and the set of N terms w are given by the following [70]:

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N_d} p(z_n | \theta) p(w_n | z_n, \beta) \tag{2}$$

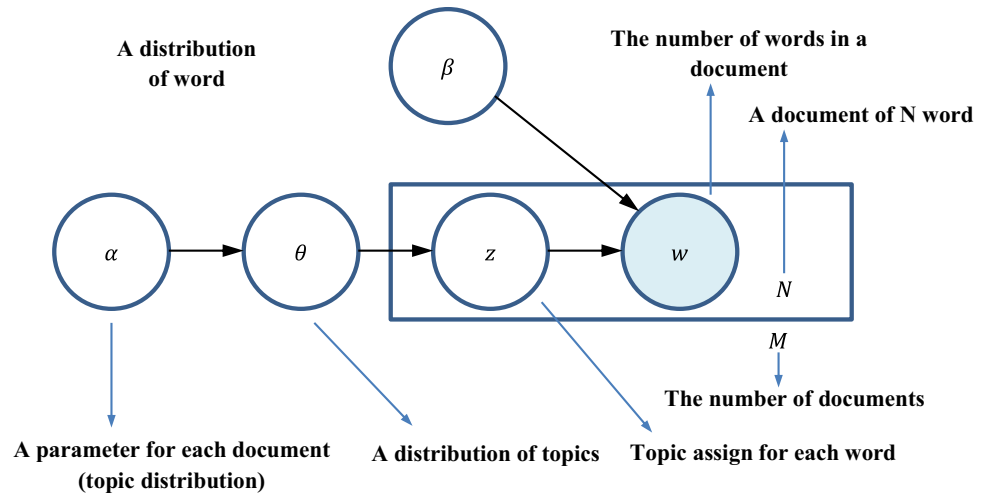
By integrating over θ , summing over z and then taking the product of the marginal probabilities of the individual documents, the probability of the entire corpus can be computed as follows:

$$p(D | \alpha, \beta) = \prod_{n=1}^{N_d} \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_n | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d \tag{3}$$

3.3 Satisfaction prediction using the SVR model

Considering that l is the number of training samples in a set of training samples $\{(x_i, y_i)\}_{i=1}^l$ with x_i as the input sample values and y_i are the corresponding model output values, a

Fig. 4 The graphical representation of LDA



regression problem is provided to estimate output y based on the input values (x):

$$y = f(x) + v \tag{4}$$

where x denotes the input variable, v denotes noise independent of x , and $f(\cdot)$ denotes an unknown nonlinear function. By introducing an alternative *loss* function (e.g., ϵ -insensitive function, Huber function, Laplace function, quadratic function), support vector machines can be applied to regression problems. It has been shown that the ϵ -insensitive function can be effective in producing sparseness in the support vectors; accordingly, the ϵ -insensitive function is used in this paper for support vector regression. The unknown function $f(x)$ in support vector regression is assumed to be expressed as follows:

$$f_{SVR}(x) = W^T \phi(x) + b \tag{5}$$

where $\phi(\cdot)$ denotes a prespecified nonlinear mapping, b denotes a bias, and w denotes a coefficient vector. Accordingly, the bias and the coefficient vector values are identified so that the following functional is minimized:

$$C \sum_{i=1}^l |y_i - f_{SVR}(x_i)|_{\epsilon} + \|w\|^2 \tag{6}$$

where $|\cdot|_{\epsilon}$ denotes an ϵ -insensitive loss function, l is the number of training samples, and C is a positive constant. The loss function is equal to zero if the difference between the observed outputs and regression is less than ϵ . The ϵ -insensitive function is calculated as follows:

$$|x|_{\epsilon} = \begin{cases} 0 & \text{if } |x| < \epsilon \\ |x| - \epsilon & \text{otherwise} \end{cases} \tag{7}$$

The following convex QP problem of the dual form of the above problem with $2l$ variables is defined to find Lagrange multipliers $\Psi = [\Psi_1, \Psi_2, \dots, \Psi_l]^T$ and $\Psi' = [\Psi'_1, \Psi'_2, \dots, \Psi'_l]^T$, which minimize the following:

$$\begin{aligned} Q(\Psi, \Psi') = & - \sum_{i=1}^l y_i (\Psi_i - \Psi'_i) + \epsilon \sum_{i=1}^l (\Psi_i + \Psi'_i) \\ & + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\Psi_i - \Psi'_i) (\Psi_j - \Psi'_j) K(x_i, x_j) \end{aligned} \tag{8}$$

subject to the following constraints:

$$\begin{aligned} \sum_{i=1}^l (\Psi_i - \Psi'_i) &= 0 \\ 0 \leq \Psi_i &\leq C, i = 1, 2, \dots, l \\ 0 \leq \Psi'_i &\leq C, i = 1, 2, \dots, l \end{aligned} \tag{9}$$

where l is the number of training samples, ϵ denotes a user-specified positive constant, $K(\cdot, \cdot)$ is called a kernel function, and $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.

Let (Ψ^*, Ψ'^*) be considered the optimal solution to the above problem. Accordingly, the $f_{SVR}(x)$, i.e., the approximating function, derived by support vector regression can be computed as follows:

$$f_{SVR}(x) = \sum_{i=1}^l (\Psi_i^* - \Psi'^*_i) K(x_i, x) + b^* \tag{10}$$

where l is the number of training samples. The Lagrange multipliers satisfy the Karush–Kuhn–Tucker conditions:

$$\Psi_i^* \cdot \Psi'^*_i = 0, i = 1, 2, \dots, l \tag{11}$$

By considering $\Upsilon_i = \Psi_i - \Psi'_i$ for all and $\Upsilon = [\Upsilon_1, \Upsilon_2, \dots, \Upsilon_l]^T$, the aim is to minimize the following problem:

$$W(\Upsilon) = - \sum_{i=1}^l y_i \Upsilon_i + \epsilon \sum_{i=1}^l |\Upsilon_i| + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \Upsilon_i \Upsilon_j k_{ij} \tag{12}$$

subject to the following constraints:

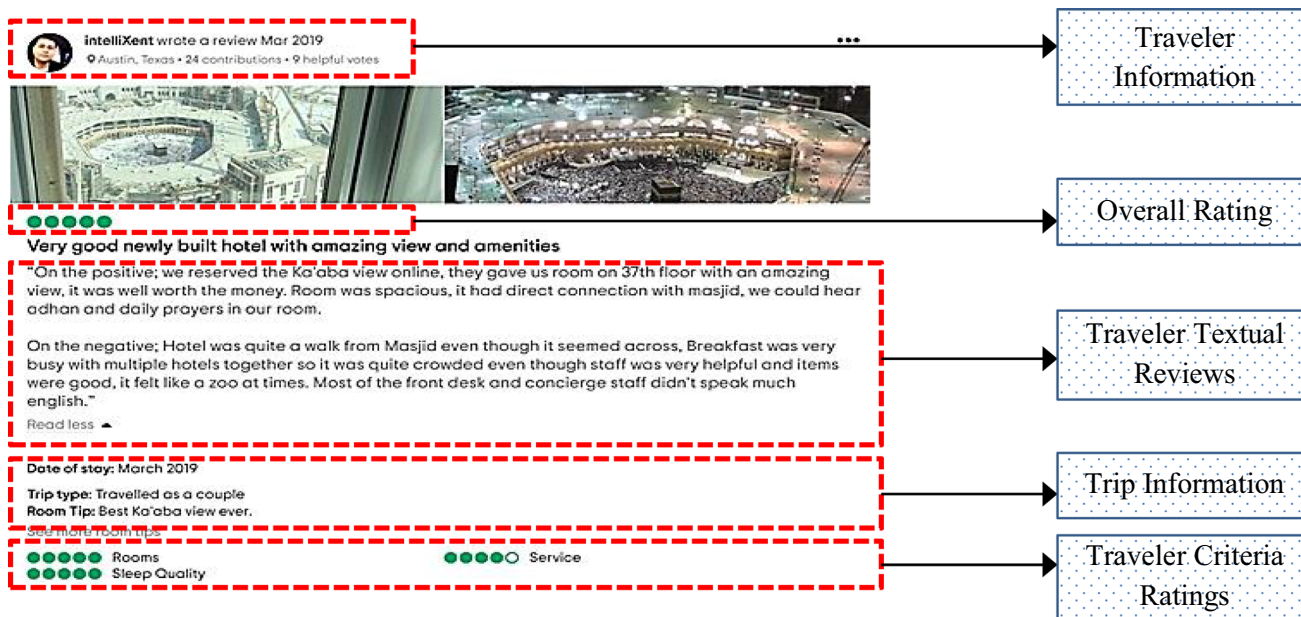


Fig. 5 Travelers’ reviews of Mecca hotels on TripAdvisor

Table 3 A sample of ratings by travelers

Hotel ID	Traveler ID	Location	Cleanliness	Service	Value	Overall rating	Travel type
Hotel 1	100	3	Null	3	4	5	Traveled with family
Hotel 2	170	4	2	4	2	4	Traveled on business
Hotel 3	30	5	5	5	4	5	Traveled with friends
Hotel 10	30	5	5	5	3	5	Traveled as a couple
–	–	–	–	–	–	–	–
Hotel 30	5	5	Null	Null	4	5	Traveled solo

Table 4 Trip types and the number of travelers

Travel type	Traveled as a couple	Traveled on business	Traveled solo	Traveled with family
	742	174	344	464

$$\sum_{i=1}^l Y_i = 0; -C \leq Y_i \leq C, i = 1, 2, l \tag{13}$$

where l is the number of training samples. By considering $Y^* = [Y_1^*, Y_2^*, Y_l^*]^T$ to be an optimal solution, $f_{SVR}(x)$ is calculated as follows:

$$f_{SVR}(x) = \sum_{i=1}^l Y_i^* K(x_i, x) + b^* \tag{14}$$

3.4 Evaluating the accuracy of the proposed models

The MAE, RMSE, and R^2 were used to describe the predictive abilities of the proposed models. The following are the definitions of these evaluation metrics:

$$MAE = \frac{1}{n} \sum_{i=1}^n |actual_i - forecast_i| \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (actual_i - forecast_i)^2}{n}} \tag{16}$$

$$R^2 = \frac{[\sum_{i=1}^n (actual_i - \overline{actual_i}) \cdot (forecast_i - \overline{forecast_i})]^2}{\sum_{i=1}^n (actual_i - \overline{actual_i})^2 \sum_{i=1}^n (forecast_i - \overline{forecast_i})^2} \tag{17}$$

where actual denotes the real overall rating, forecast denotes the predicted overall rating, \overline{actual} denotes the mean of the observed values, $\overline{forecast}$ denotes the mean of the predicted values, and n is the number of samples.

Table 5 Information on hotel ratings based on trip types and performance criteria

Criteria	Ratings	Travel type			
		Traveled as a couple	Traveled on business	Traveled solo	Traveled with family
Location	1	8	3	4	7
	2	44	15	32	22
	3	164	36	80	106
	4	197	43	87	137
	5	329	77	141	192
Cleanliness	1	18	6	11	15
	2	108	26	50	51
	3	153	33	76	108
	4	190	38	92	110
	5	273	71	115	180
Service	1	33	11	21	24
	2	69	14	19	30
	3	162	40	80	109
	4	233	44	109	153
	5	245	65	115	148
Value	1	47	12	17	29
	2	70	13	19	36
	3	47	16	21	24
	4	253	60	127	176
	5	325	73	160	199

Table 6 Silhouette coefficient value segments obtained by *k*-means

Number of segments	Silhouette coefficient	Number of segments	Silhouette coefficient
2	0.6634	8	0.8441
3	0.6998	9	0.7837
4	0.7231	10	0.7647
5	0.8216	11	0.7894
6	0.8969	12	0.7722
7	0.7492	13	0.7621

Table 7 Cluster centroids and size

Attribute	Segment 1 (237)	Segment 2 (465)	Segment 3 (344)	Segment 4 (326)	Segment 5 (184)	Segment 6 (168)
Location	4.578059	4.746237	3.095930	4.407975	4.168478	2.357143
Cleanliness	4.451477	4.733333	4.130814	2.472393	3.222826	2.791667
Service	2.497890	4.660215	4.148256	4.036810	3.385870	2.511905
Value	4.392405	4.587097	4.415698	4.496933	1.581522	3.000000

4 Method evaluation

In this section, we provide the data collection, data analysis and method evaluation. The results are presented and discussed.

4.1 Data collection

This research collected data from TripAdvisor. TripAdvisor is a trusted source of information for travelers and has been providing quality reviews since 1995 [97]. All

Table 8 Overall ratings and trip types for the six segments

Overall rating	Segments						Sum
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	All segments
1	0	1	0	0	5	42	48
2	0	23	26	25	27	77	178
3	28	99	62	62	75	41	367
4	185	218	189	159	77	8	836
5	24	124	67	80	0	0	295

Trip types	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	All segments
Traveled as a couple	114	199	139	136	86	68	742
Traveled on business	24	48	29	33	17	23	174
Traveled solo	40	91	80	67	28	38	344
Traveled with family	59	127	96	90	53	39	464

reviews are passed through a technology-based moderation process and are further screened by the content moderation team. Hence, it was found that TripAdvisor might be a trusted source to investigate customer satisfaction with hotels in Mecca, Saudi Arabia. The hotels in Mecca, Saudi Arabia, that were registered on TripAdvisor were randomly selected. In addition, we selected 4- and 5-star Mecca hotels for data collection. The online reviews were in the form of numerical and textual reviews with other information on the hotels, such as hotel stars, the travel type, the date of the stay, and the date of the review. To better present the impacts of performance quality factors on customer satisfaction, the reviews that only provided overall ratings without criteria ratings were removed from the dataset. In Fig. 5, we provide an example of travelers' reviews of Mecca hotels on TripAdvisor. Two types of reviews are provided by travelers. They are in the form of textual reviews and numerical ratings for criteria and overall ratings in the range of +1 to +5. When crawling the data from TripAdvisor, it was found that many travelers provided ratings solely for the overall ratings; thus, we removed these records from the datasets, as they could impact the accuracy of the overall rating prediction through a set of criteria ratings. Doing so could improve the quality of the collected data to achieve the objectives of this research. In total, 1724 records were collected from 30 hotels in Mecca, Saudi Arabia. These records of ratings and reviews were further analyzed by the proposed method. A sample of ratings by travelers is provided in Table 3. As shown in Table 3, location, cleanliness, service, and value were the main criteria for overall rating assessment. In addition, traveled solo, traveled with friends, traveled with family, traveled on business and traveled as a couple were the major trip types of travelers in hotels in Mecca, Saudi

Arabia. The trip types and the number of travelers are presented in Table 4. In Table 5, information on hotel ratings based on trip types and performance criteria is presented.

4.2 Data segmentation and satisfaction analysis

The 1724 records of travelers' ratings were used in the *k*-means algorithm for data segmentation. Different *k* values were considered for data segmentation. The optimal number of clusters was selected by the silhouette coefficient approach. The results are shown in Table 6 for different *k* values. It is found that the *k*-means algorithm provides the best segments for *k* = 6. In fact, when *k* = 6, the silhouette coefficient is higher than the other number of clusters. The value of the silhouette coefficient in six segments is approximately 0.8969. It is shown that there is a direct relationship between the quality of clustering and the accuracy of prediction learning techniques [98]. Hence, 6 segments were selected for further evaluation by SVR.

The cluster centroids based on the performance criteria are presented in Table 7. High ratings were provided for location, cleanliness, and value in Segment 1. In Segment 2, the ratings were high for all criteria. In Segment 3, cleanliness, service, and value received high ratings from travelers. In Segment 4, location, service and value received high ratings. In Segment 5 and Segment 6, location and value obtained a high level of ratings by travelers in these groups. Overall, the results of Table 5 show that the majority of travelers are in Segment 2, Segment 3, and Segment 4. In Table 6, we present information on the overall ratings and trip types for the six segments. It is clearly found that in Segments 1–4, travelers' overall ratings were mainly in the range of 3 to 5; however, in

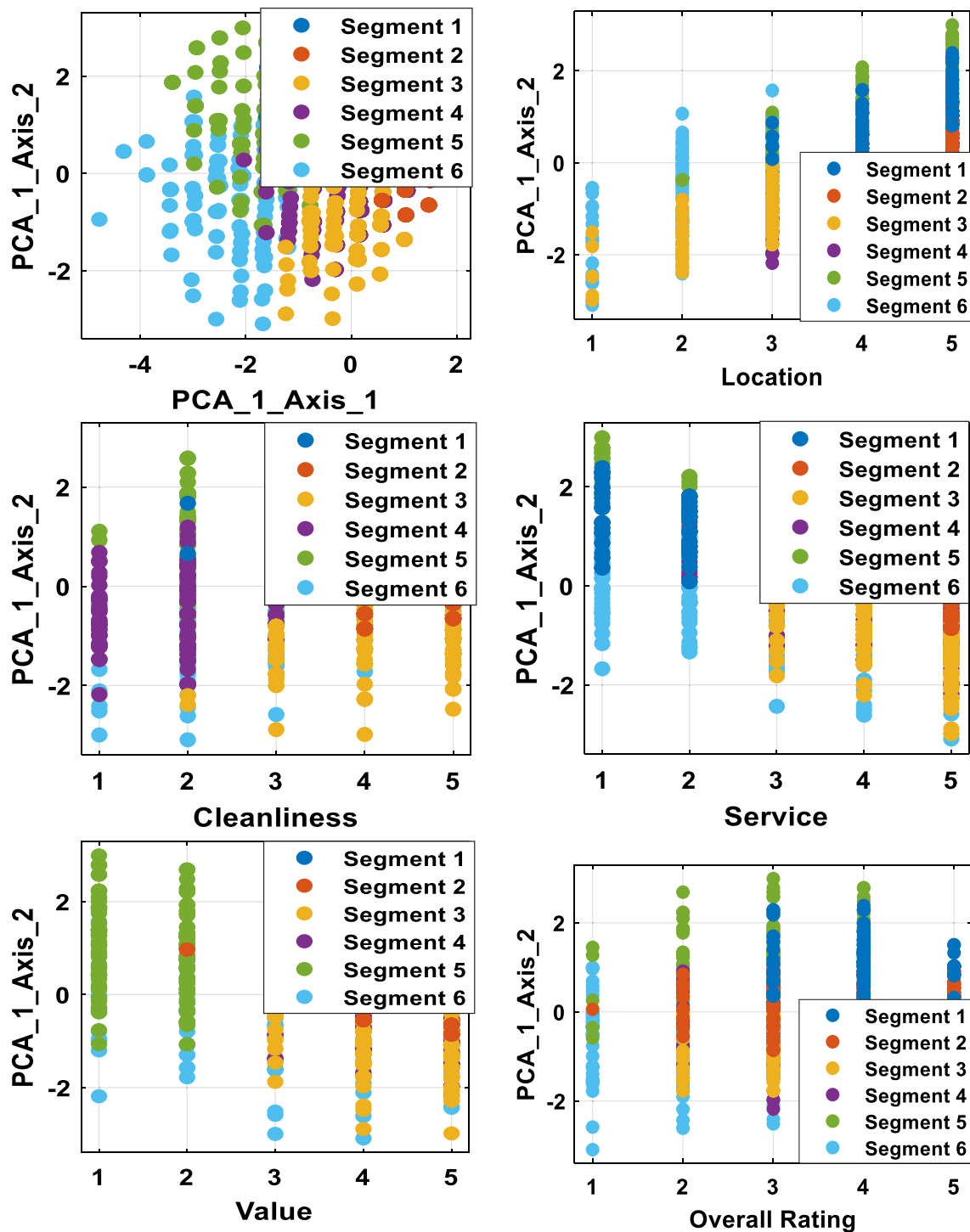


Fig. 6 Segment visualization

Segment 5, the overall ratings were more in the range of 3 to 4. In addition, in Segment 6, the overall ratings show that travelers provided ratings mainly in the range of 1 to 3. Furthermore, the results in Table 8 show that the majority of travelers in all segments were in the traveled as a couple

group. The segments and the criteria for each segment are visualized in Fig. 6.

We also performed textual data analysis on each segment using LDA. The online traveler reviews in the form of textual data were analyzed to discover the main dimensions of satisfaction for each segment. The results are

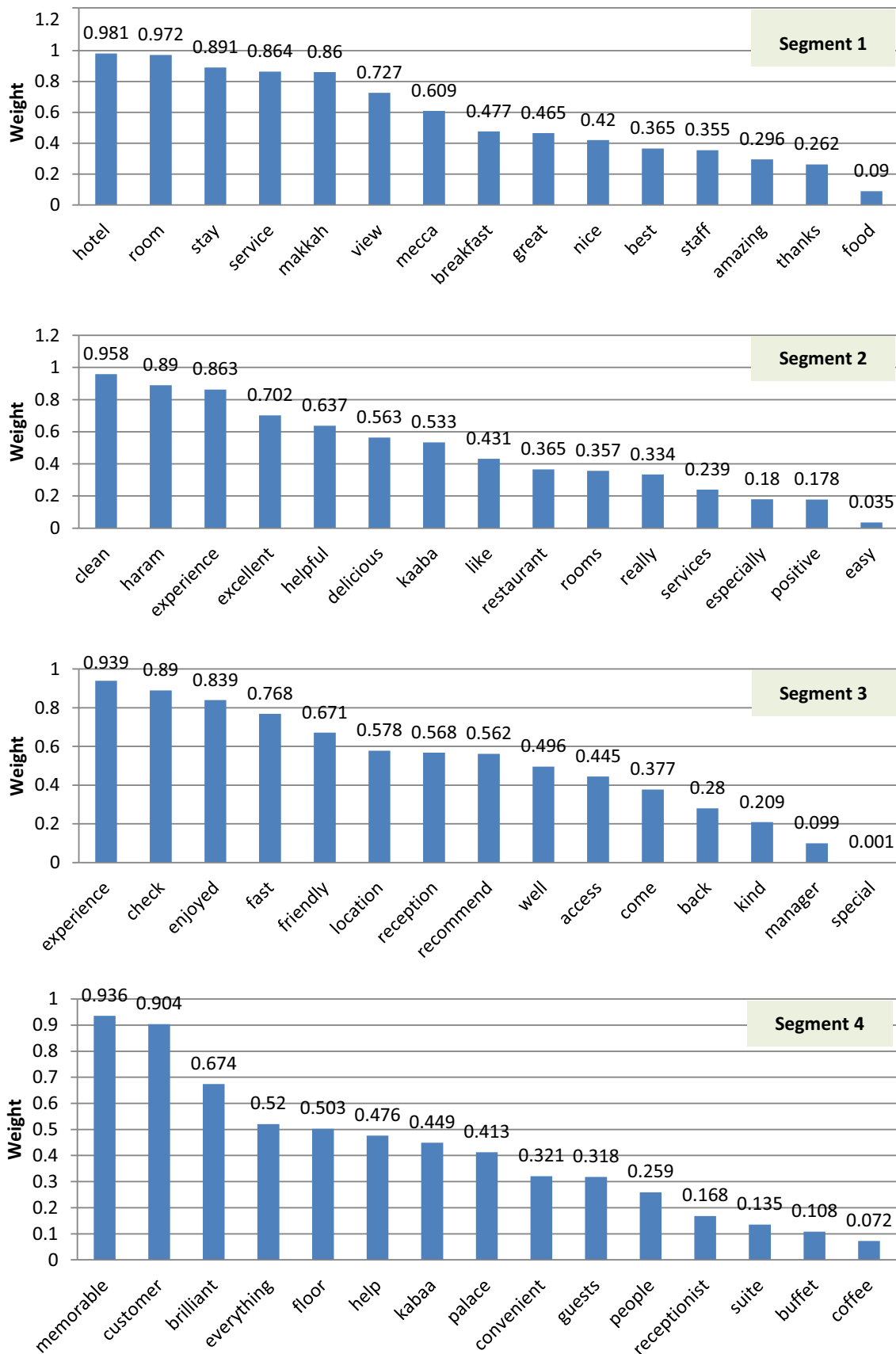


Fig. 7 Top 15 satisfaction dimensions for the six segments

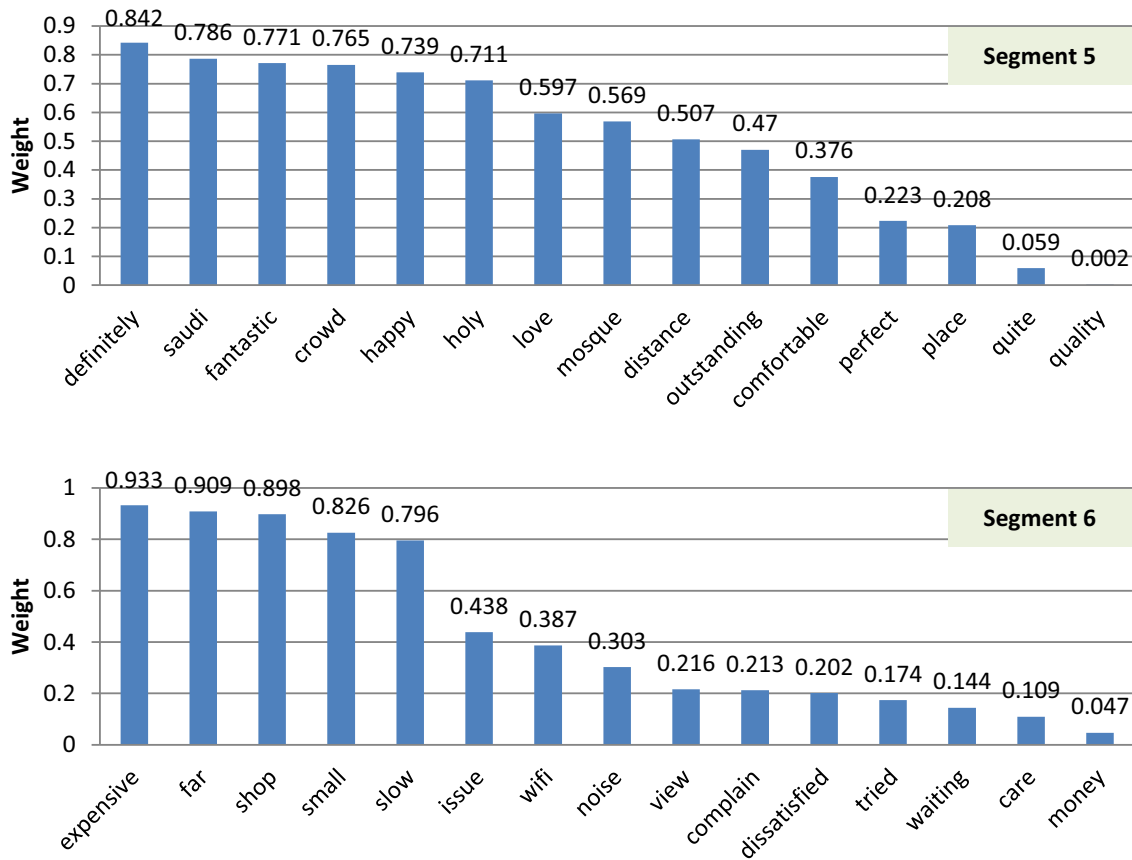


Fig. 7 continued

Table 9 SVR parameters for epsilon-SVR

SVR parameters	Parameter value	
Epsilon	0.05	0.07
Kernel type	RBF	LINEAR
Degree (poly)	1	1
Tolerance of termination criteria	0.002	0.003
C (Complexity Cost)	1	2
Use shrinking heuristics	1	3
No. support vector	67	78

shown in Fig. 7 for the top 15 keywords. The results of textual data analysis show that travelers are satisfied with the quality of services in hotels in Mecca, Saudi Arabia, for 5 segments. Segment 6, in which the satisfaction level is low, is the exception. The results show that travelers are mainly dissatisfied with the price of hotels, the slow service in hotel preparation, room size, wifi and internet access, and their hotel’s distance from shopping centers.

4.3 Satisfaction prediction using SVR

After traveler segmentation through the use of online reviews, it was important to predict customer satisfaction using a set of performance criteria for each segment. To do so, we implemented SVR on each segment and predicted customer satisfaction based on the service quality on four performance criteria: location, cleanliness, service, and value. To better perform the prediction, we optimized the SVR algorithm using the SMO solver algorithm [99]. In fact, we adopted the SMO solver algorithm to train the SVR models. In addition, we adopted two different kernel functions, RBF and linear functions [100, 101], to find the best SVR models based on the RMSE, MAE and R² metrics. We used tenfold cross-validation to train the SVR models. The parameter specification of the models of the support vector regression is shown in Table 9.

Based on the above SVR parameters, we trained the models on a training set and then provided the results for the test set in each segment for the RMSE, MAE and R². The results are shown in Fig. 8 for the RMSE and MAE for the RBF and linear kernel functions. Based on the MAE and RMSE results, SVR effectively constructed the prediction models for the six segments based on low RMSE

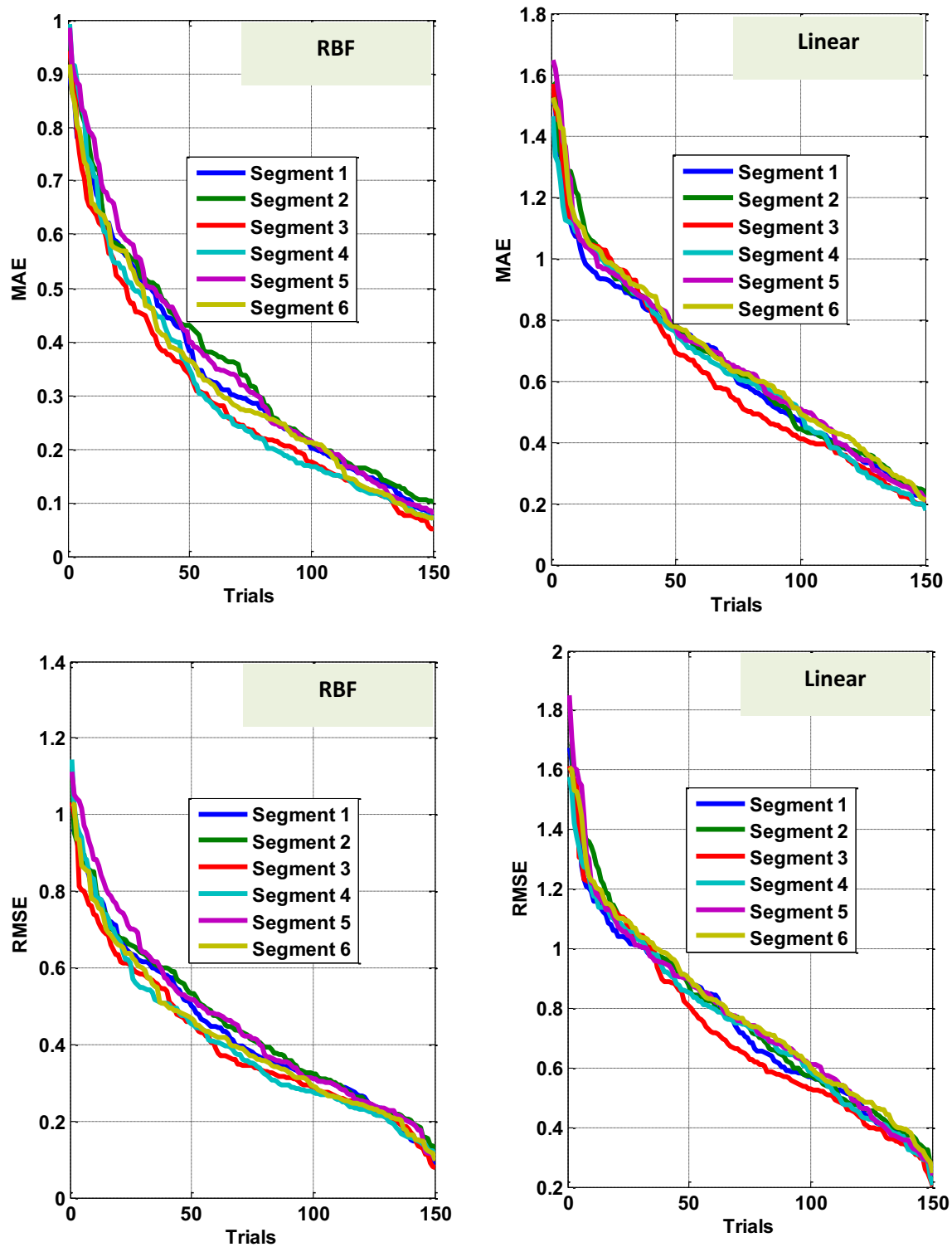


Fig. 8 The MAE and RMSE in 150 trials using the training set

and MAE values. The results in Fig. 8 demonstrate that the RBF kernel function provided better RMSE and MAE values for the six segments in relation to the linear kernel.

The R^2 results for the six segments are presented in Fig. 9. In addition, the average evaluation results of the method based on the test sets for the six segments are presented in Table 10. In this table, we also compare our

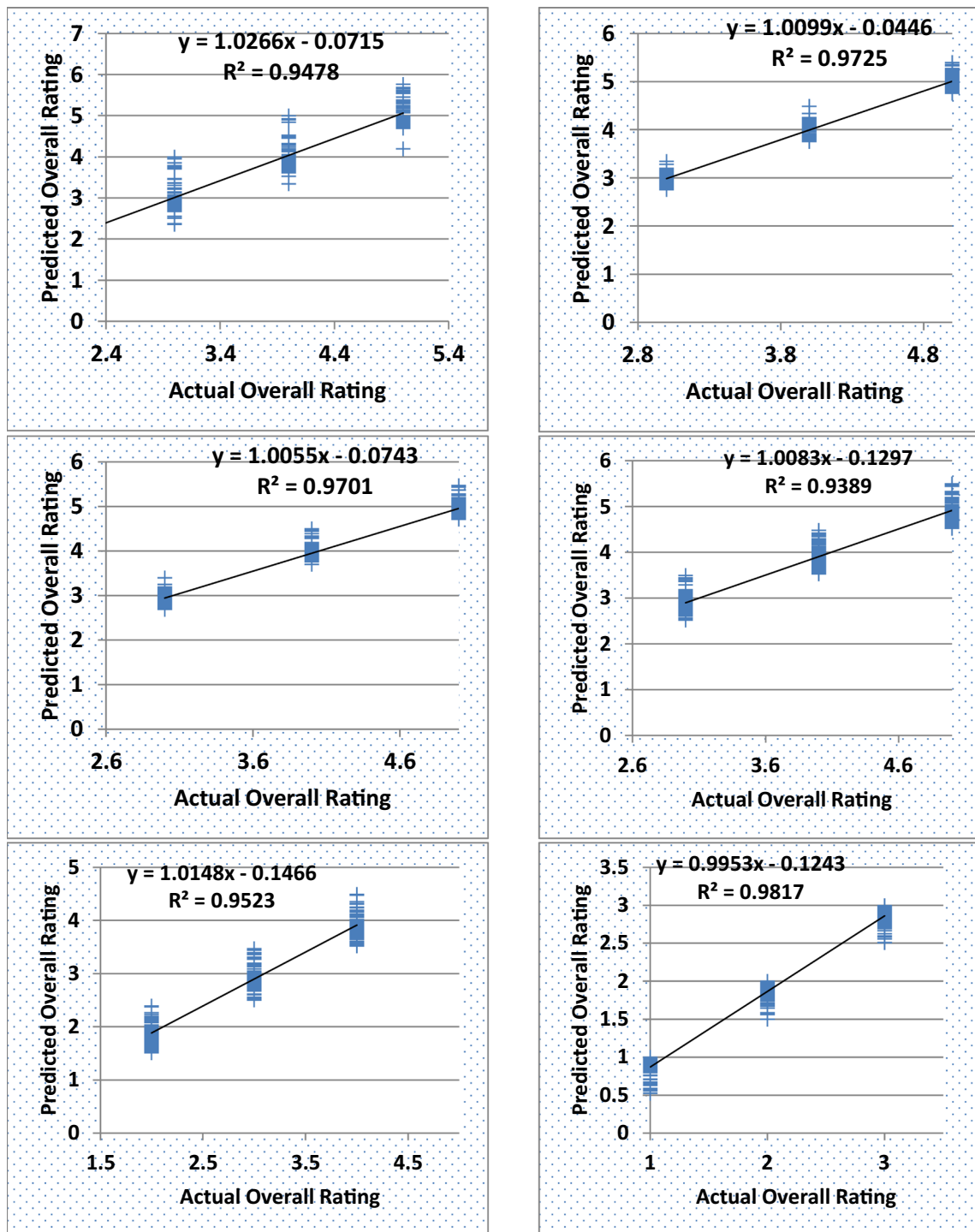


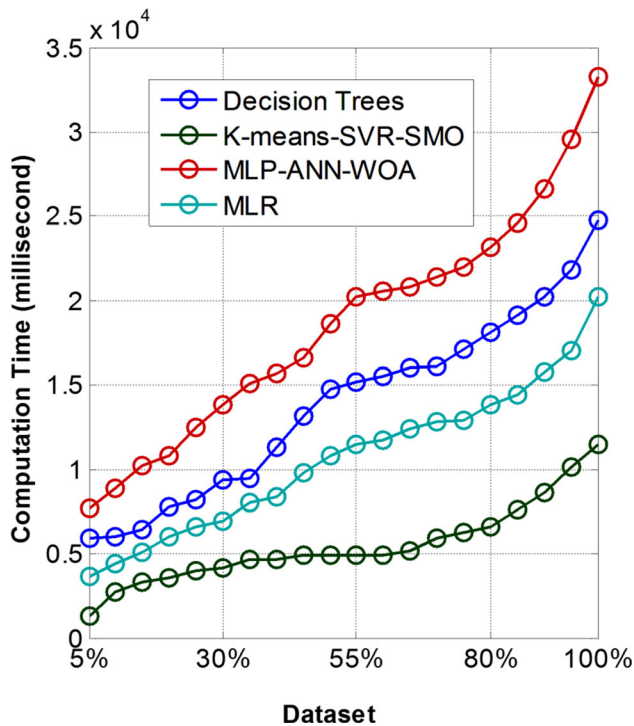
Fig. 9 R^2 results for the six segments

results with those of other prediction learning techniques. Specifically, we compare our results with the multilayer perceptron artificial neural network (MLP-ANN) [102], multiple linear regression (MLR), and decision trees [103]. In this comparison, MLP is optimized with the whale optimization algorithm (WOA) [104, 105].

The R^2 results for the six segments show that the method that used clustering and SVR with SMO optimization provided accurate prediction results for the overall rating. In addition, on average, *k*-means-SVR-SMO (RMSE = 0.1042; MAE = 0.2823; $R^2 = 0.9526$) was more accurate in the prediction of overall ratings than MLP-ANN-WOA

Table 10 Method evaluation results

Method	RMSE	MAE	R^2
K-means-SVR-SMO	0.1042	0.2823	0.9526
MLP-ANN-WOA	0.6741	0.8213	0.8278
Decision Trees	0.8321	1.1024	0.7925
MLR	1.2132	1.3751	0.7324

**Fig. 10** Computation time of different methods

(RMSE = 0.6741; MAE = 0.8213; $R^2 = 0.8278$), decision trees (RMSE = 0.8321; MAE = 1.1024; $R^2 = 0.7925$), and MLR (RMSE = 1.2132; MAE = 1.3751; $R^2 = 0.7324$). In this comparison, MLR, decision trees and MLP-ANN-WOA were trained on the whole dataset with tenfold cross-validation, and MLP-ANN-WOA was trained for 150 epochs.

We also aimed to compare the computation time of the proposed method with that of other prediction learning techniques. Accordingly, at each time point, 5% of the data were considered to test the prediction models, and the computation time was calculated. The results are presented in Fig. 10. The results show that the computation time of *k*-means-SVR-SMO was less than that of MLP-ANN-WOA, decision trees and MLR. This finding indicates the effectiveness of integrating the clustering approach with SVR in reducing the prediction computation time. Note that all implementations were performed in MATLAB software on

an Intel(R) Core(TM) i7-6700HQ CPU clocked at 2.60 GHz and with 8 GB RAM.

5 Discussion and research implications

Tourism has long been acknowledged as one of the world's most important socioeconomic factors [106, 107]. Research and practice have been more prevalent in recent decades as a result of the expansion of online review sites. In the e-commerce and big data era, online customer reviews of hotels have enormous business value. Online textual reviews have an open structure in which customers can better present their preferences, feedback, and opinions on the quality of services. In fact, online textual reviews are a direct measure of a customer's view of a product or service, and in this way, they can more fully reflect the consumption experience and the customer's perception. In machine learning, they can be useful for revealing customer satisfaction. This has led to the growth of studies on online reviews and the application of new analytical methods. Online review analysis has been effective for customer preference learning. Traveler segmentation has recently been an important topic in the tourism and hospitality context [108–110]. Preference learning has a significant impact on effective customer segmentation. Machine learning has played an important role in business, customer preference learning, and segmentation [111, 112]. Data availability, the speed and simplicity of data collection, and nonintrusiveness with regard to human participants are only a few of the benefits of performing customer segmentation through the use of online reviews. Accordingly, there have been several attempts to develop new methods in this context [36, 113, 114]. This study also tried to develop a new method for traveler segmentation with online customer reviews.

Our major contribution to this research was the development of a new framework for collecting and processing big social data, extracting the dimensions of customer satisfaction, predicting customer satisfaction, and identifying customer preferences for several segments. The method was effectively evaluated based on online traveler reviews. We collected 1724 online reviews of 30 hotels in Mecca, Saudi Arabia. To collect such a large amount of data for customer satisfaction analysis, the use of a questionnaire survey could be difficult. The proposed method could be suitable for large dataset analysis, as it includes a clustering approach before predicting customer satisfaction. In fact, the segmentation stage could play an important role in the effectiveness of the proposed method. This can be verified by comparing the results of our method with those of methods that are not based on clustering approaches. In addition, when large datasets of online reviews are

available, clustering may be beneficial for decreasing the computation time of predicting customer preferences and satisfaction. Therefore, in future studies, it could be interesting to improve the segmentation part of the proposed method for large datasets with additional performance criteria.

The results of the customer satisfaction analysis through the use of the segmentation approach were interesting. We found six main segments from the collected data. Each segment provided valuable information on travelers' preferences. It was revealed that location, cleanliness, and value were more important for Segment 1. In Segment 2, the rating analysis showed that all criteria were important to travelers. In Segment 3, cleanliness, service, and value were the most important performance criteria for Mecca hotels. In Segment 4, location, service and value were the most important performance criteria; in Segment 5 and Segment 6, they were location and value. From textual review analysis, it was found that the most important dimensions of satisfaction were hotel, room, stay, service, makkah, view, mecca, breakfast, great, nice, best, staff, amazing, thanks and food for Segment 1. In Segment 2, clean, haram, experience, excellent, helpful, delicious, kaaba, like, restaurant, rooms, really, services, especially, positive, and easy were the most important dimensions of satisfaction. In Segment 3, the most important dimensions of satisfaction were experience, check, enjoyed, fast, friendly, location, reception, recommend, well, access, come, back, kind, manager and special. In Segment 4, they were memorable, customer, brilliant, everything, floor, help, kabaa, palace, convenient, guests, people, receptionist, suite, buffet and coffee. In Segment 5, the most important dimensions of satisfaction were definitely, saudi, fantastic, crowd, happy, holy, love, mosque, distance, outstanding, comfortable, perfect, place, quite and quality. In Segment 6, they were expensive, far, shop, small, slow, issue, wifi, noise, view, complain, dissatisfied, tried, waiting, care and money. Overall, it was found that the results of numerical ratings are in line with the textual review analysis. The results of textual review analysis showed that of the six segments discovered by *k*-means, only the last segment includes travelers with low satisfaction levels, which was also identified from the numerical rating analysis.

Overall, the analysis of online reviews through the implementation of new machine learning approaches is an important task. It is shown that the results of online reviews can effectively help hotel managers better detect traveler demands. In addition, hotel managers can be aware of their level of quality service and limitations in regard to different aspects of their hotels' performance criteria. Online reviews present a significant eWOM effect, which affects future customers' booking selections. The outcomes of this

study can serve as a motivator for hotel managers to extract additional attributes from customer textual reviews and to conduct in-depth investigations into customers' decision-making behavior through their online reviews and their relationship with satisfaction and purchase intention. As online reviews include both negative and positive feedback on the quality of services, responding quickly and efficiently to negative online customer evaluations is an effective way to retain consumers and conduct service recovery steps. In addition, hotel managers' involvement in online reviews can significantly impact customers' perception of their hotels' services and increase customers' trust to obtain future purchases.

6 Conclusion

This research developed a new method of customer segmentation using support vector regression with sequential minimal optimization, latent Dirichlet allocation (LDA) and *k*-means approaches. The data on travelers' online reviews of Mecca hotels were collected from TripAdvisor. The data were segmented using *k*-means, and traveler satisfaction was revealed for each segment. The LDA technique was used for textual data analysis and to discover the dimensions of satisfaction for each segment. Finally, SVR-SMO was applied in each segment to predict customer satisfaction (overall ratings) based on several performance quality factors. The effectiveness of the proposed method was measured using the RMSE, MAE and R^2 . The results of the evaluations were then compared with those based on the ANN, MLR and decision tree approaches. The results revealed that the predictive accuracy of the method that used *k*-means and SVR-SMO was higher than that of the other methods. This research has some limitations that need to be addressed in future studies. This study used TripAdvisor for data collection. Although the results were provided based on online reviews of hotels in Mecca, Saudi Arabia, the results could be further verified based on additional data collected by a questionnaire survey. In the case of the proposed method, the R^2 results show that the predictive accuracy of the method could still be improved. As standard SVR was used for satisfaction prediction, the use of online SVR could be interesting research for enhancing the predictive accuracy and customer satisfaction analysis. In addition, the use of other clustering approaches for traveler segmentation is suggested. Furthermore, this research did not investigate customer satisfaction with each travel type through hotel performance criteria. Accordingly, it could be interesting to find the relationships between hotel performance criteria and customer satisfaction for each travel group. Doing so may reveal customers' preferences for different trip types based

on their experience with hotel services. Moreover, the silhouette coefficient was used in this study to find the optimal number of clusters based on the collected data. Accordingly, other approaches to evaluating clustering quality could be investigated.

Data availability All the data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

- Halim YT, Halim HT (2013) Guest Satisfaction and Hotel Profitability in Egypt. *Journal of Association of Arab Universities for Tourism and Hospitality*, 10(1).
- Laškarin Ažić M, Dlačić J, Suštar N (2020) Loyalty trends and issues in tourism research. *Tourism and Hospitality Manag* 26(1):133–155
- Chang Y-C, Ku C-H, Chen C-H (2019) Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *Int J Inf Manage* 48:263–279
- Tran LTT, Ly PTM, Le LT (2019) Hotel choice: a closer look at demographics and online ratings. *Int J Hosp Manag* 82:13–21
- Barsky JD (1992) Customer satisfaction in the hotel industry: Meaning and measurement. *Hospitality Res J* 16(1):51–73
- Ladhari R, Michaud M (2015) eWOM effects on hotel booking intentions, attitudes, trust, and website perceptions. *Int J Hosp Manag* 46:36–45
- Ahani A et al (2019) Revealing customers' satisfaction and preferences through online review analysis: the case of Canary Islands hotels. *J Retail Consum Serv* 51:331–343
- Kandampully J, Zhang TC, Jaakkola E (2018) Customer experience management in hospitality: A literature synthesis, new understanding and research agenda. *Int J Contemporary Hospitality Manag*
- Zhang Z, Ye Q, Law R (2011) Determinants of hotel room price: an exploration of travelers' hierarchy of accommodation needs. *Int J Contemporary Hospitality Manag*
- Kim RY (2019) Does national culture explain consumers' reliance on online reviews? Cross-cultural variations in the effect of online review ratings on consumer choice. *Electron Commer Res Appl* 37:100878
- Berezina K et al (2016) Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *J Hosp Market Manag* 25(1):1–24
- Kuhzady S, Ghasemi V (2019) Factors influencing customers' satisfaction and dissatisfaction with hotels: a text-mining approach. *Tour Anal* 24(1):69–79
- Ye Q et al (2011) The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Comput Hum Behav* 27(2):634–639
- Gao S et al (2018) Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *Int J Hosp Manag* 71:19–32
- Xu X, Li Y (2016) The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: a text mining approach. *Int J Hosp Manag* 55:57–69
- Xiang Z et al (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? *Int J Hosp Manag* 44:120–130
- Ahani A et al (2019) Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews. *Int J Hosp Manag* 80:52–77
- Narangajavana Kaosiri Y et al (2019) User-generated content sources in social media: A new approach to explore tourist satisfaction. *J Travel Res* 58(2):253–265
- Chan KY, Kwong C, Kremer GE (2020) Predicting customer satisfaction based on online reviews and hybrid ensemble genetic programming algorithms. *Eng Appl Artif Intell* 95:103902
- Bi J-W et al (2019) Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. *Int J Prod Res* 57(22):7068–7088
- Sánchez-Franco MJ, Navarro-García A, Rondán-Cataluña FJ (2019) A naive Bayes strategy for classifying customer satisfaction: a study based on online reviews of hospitality services. *J Bus Res* 101:499–506
- Green PE (1977) A new approach to market segmentation. *Bus Horiz* 20(1):61–73
- Füller J, Matzler K (2008) Customer delight and market segmentation: An application of the three-factor theory of customer satisfaction on life style groups. *Tour Manage* 29(1):116–126
- Dolnicar S, Grün B, Leisch F (2018) Market segmentation analysis: Understanding it, doing it, and making it useful: Springer Nature.
- Bloom JZ (2004) Tourist market segmentation with linear and non-linear techniques. *Tour Manage* 25(6):723–733
- Dolnicar S (2002) A review of data-driven market segmentation in tourism. *J Travel Tour Mark* 12(1):1–22
- Gonzalez AM, Bello (2002) The construct "lifestyle" in market segmentation: The behaviour of tourist consumers. *Eur J Market*
- Kuo H-C, Nakhata C (2019) The impact of electronic word-of-mouth on customer satisfaction. *J Market Theory and Practice* 27(3):331–348
- Liu H et al (2021) Social sharing of consumption emotion in electronic word of mouth (eWOM): a cross-media perspective. *J Bus Res* 132:208–220
- Huete-Alcocer N (2017) A literature review of word of mouth and electronic word of mouth: implications for consumer behavior. *Front Psychol* 8:1256
- Khorsand R, Rafiee M, Kayvanfar V (2020) Insights into TripAdvisor's online reviews: the case of Tehran's hotels. *Tourism Manag Perspect* 34:100673
- Nilashi M et al (2021) Travellers decision making through preferences learning: a case on Malaysian spa hotels in TripAdvisor. *Comput Ind Eng* 158:107348
- Nilashi M et al (2018) Travelers decision making using online review in social network sites: a case on TripAdvisor. *J Comput Sci* 28:168–179
- Arenas-Márquez FJ, Martínez-Torres R, Toral S (2021) Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor. *Inf Process Manage* 58(5):102645
- Fernandes E et al (2021) A data-driven approach to measure restaurant performance by combining online reviews with historical sales data. *Int J Hosp Manag* 94:102830
- Taecharungroj V, Mathayomchan B (2019) Analysing TripAdvisor reviews of tourist attractions in Phuket. *Thailand Tourism Manag* 75:550–568

37. Gebbels M, McIntosh A, Harkison T (2021) Fine-dining in prisons: Online TripAdvisor reviews of The Clink training restaurants. *Int J Hosp Manag* 95:102937
38. Bigne E et al (2021) What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations. *J Destin Mark Manag* 20:100570
39. Song Y et al (2021) Investigating sense of place of the Las Vegas Strip using online reviews and machine learning approaches. *Landsc Urban Plan* 205:103956
40. Borges-Tiago MT et al (2021) Differences between TripAdvisor and Booking.com in branding co-creation. *J Bus Res* 123:380–388
41. Zhang C et al (2021) An online reviews-driven method for the prioritization of improvements in hotel services. *Tour Manage* 87:104382
42. Korfiatis N et al (2019) Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Syst Appl* 116:472–486
43. Kotler P et al. (2014) *Marketing management 14/e*. 2014: Pearson
44. Elrod C et al (2015) Empirical study utilizing QFD to develop an international marketing strategy. *Sustainability* 7(8):10756–10769
45. Rahim MA et al (2021) RFM-based repurchase behavior for customer classification and segmentation. *J Retail Consum Serv* 61:102566
46. Huseynov F, Yıldırım SÖ (2017) Behavioural segmentation analysis of online consumer audience in Turkey by using real e-commerce transaction data. *Int J Econom Bus Res* 14(1):12–28
47. Wang O, Somogyi S (2019) Consumer adoption of sustainable shellfish in China: effects of psychological factors and segmentation. *J Clean Prod* 206:966–975
48. Yadegaridehkordi E et al (2021) Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. *Technol Soc* 65:101528
49. Alkhayrat M, Aljnidi M, Aljoumaa K (2020) A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *J Big Data* 7(1):1–23
50. Sivaguru M, Punniyamoorthy M (2020) Modified dynamic fuzzy c-means clustering algorithm—Application in dynamic customer segmentation. *Appl Intell* 50(6):1922–1942
51. Sun Z-H et al (2021) GPHC: A heuristic clustering method to customer segmentation. *Appl Soft Comput* 111:107677
52. Wu J et al. (2020) An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Math Problem Eng* 2020.
53. Wu T, Liu X (2020) A dynamic interval type-2 fuzzy customer segmentation model and its application in E-commerce. *Appl Soft Comput* 94:106366
54. Akar E (2021) Customers' online purchase intentions and customer segmentation during the period of COVID-19 pandemic. *J Internet Commerce* 2021:1–31
55. Vohra R et al (2020) Using self organizing maps and K means clustering based on RFM model for customer segmentation in the online retail business. In *International Conference on Intelligent Computing*. Springer.
56. Ali A (2018) Travel and tourism: growth potentials and contribution to the GDP of Saudi Arabia. *Probl Perspect Manag* 16(1):417–427
57. Musa EYM (2021) The impact of tourism in the kingdom of Saudi Arabia on GDP, (2005–2017: An analytical approach). *Global J Econom Bus* 10(2):458–462
58. Banerjee S, Chua AY (2016) In search of patterns among travellers' hotel ratings in TripAdvisor. *Tour Manage* 53:125–131
59. Cenni I, Goethals P (2017) Negative hotel reviews on TripAdvisor: a cross-linguistic analysis. *Dis Context & Med* 16:22–30
60. Liu Y et al (2017) Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. *Tour Manage* 59:554–563
61. Kanungo T et al (2002) An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans Pattern Anal Mach Intell* 24(7):881–892
62. Celebi ME, Kingravi HA, Vela PA (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst Appl* 40(1):200–210
63. Khan SS, Ahmad A (2004) Cluster center initialization algorithm for K-means clustering. *Pattern Recogn Lett* 25(11):1293–1302
64. Henriques J et al (2020) Combining k-means and xgboost models for anomaly detection using log datasets. *Electronics* 9(7):1164
65. Ma G et al (2015) An enriched K-means clustering method for grouping fractures with meliorated initial centers. *Arab J Geosci* 8(4):1881–1893
66. Malinen MI, Mariescu-Istodor R, Fränti P (2014) K-means*: Clustering by gradual data transformation. *Pattern Recogn* 47(10):3376–3386
67. Zhou HB, Gao JT (2014) Automatic method for determining cluster number based on silhouette coefficient. In: *Advanced Materials Research*. 2014. Trans Tech Publ.
68. Addagarla SK, Amalanathan A (2020) Probabilistic unsupervised machine learning approach for a similar image recommender system for E-commerce. *Symmetry* 12(11):1783
69. Jelodar H et al (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools and Appl* 78(11):15169–15211
70. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
71. Putri I, Kusumaningrum R (2017) Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. In: *Journal of Physics: Conference Series*. 2017. IOP Publishing.
72. DiMaggio P, Nag M, Blei D (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics* 41(6):570–606
73. Castro-Neto M et al (2009) Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Syst Appl* 36(3):6164–6173
74. Hong W-C et al (2011) Forecasting urban traffic flow by SVR with continuous ACO. *Appl Math Model* 35(3):1282–1291
75. Cao Y et al (2016) Failure prognosis for electro-mechanical actuators based on improved SMO-SVR method. In: *2016 IEEE Chinese Guidance, Navigation and Control Conference (CGNCC)*. 2016. IEEE.
76. Yang J-F et al (2007) SMO algorithm applied in time series model building and forecast. in *2007 International Conference on Machine Learning and Cybernetics*. 2007. IEEE.
77. Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci Model Develop* 7(3):1247–1250
78. Jere S et al (2019) Forecasting annual international tourist arrivals in zambia using holt-winters exponential smoothing. *Open J Stat* 9(2):258–267

79. Nilashi M et al (2022) Knowledge discovery for course choice decision in massive open online courses using machine learning approaches. *Exp Syst Appl* 199:117092
80. Nilashi M., et al. (2022) Customer satisfaction analysis and preference prediction in historic sites through electronic word of mouth. *Neural Comput Appl*, p. 1–15.
81. Nilashi M et al (2020) Remote tracking of Parkinson's disease progression using ensembles of deep belief network and self-organizing map. *Expert Syst Appl* 159:113562
82. Martínez RM, Galván MO, Lafuente AMG (2014) Public policies and tourism marketing. an analysis of the competitiveness on tourism in Morelia, Mexico and Alcala de Henares Spain. *Procedia-Soc Behav Sci* 148:146–152
83. Saroyo P, Mulyati GT (2015) Analysis of prospect of agro-tourism attractiveness based on location characteristics. *Agriculture and Agricultural Sci Proc* 3:72–77
84. Nilashi M et al (2021) Big social data and customer decision making in vegetarian restaurants: a combined machine learning method. *J Retail Consum Serv* 62:102630
85. Zhang P (1993) Model selection via multifold cross validation. *Ann Statistics* 1993:299–313
86. Borra S, Di Ciaccio A (2010) Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Comput Statistics & Data Anal* 54:2976–2989
87. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967. Oakland, CA, USA
88. Sardar TH, Ansari Z (2018) An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm. *Future Comput Informat J* 3(2):200–209
89. Shahrivari S, Jalili S (2016) Single-pass and linear-time k-means clustering based on MapReduce. *Inf Syst* 60:1–12
90. Cai Y, Tang C (2021) Privacy of outsourced two-party k-means clustering. *Concurrency and Comput: Practice and Exp* 33(8):e5473
91. Estlick M et al. (2001) Algorithmic transformations in the implementation of k-means clustering on reconfigurable hardware. In: *Proceedings of the 2001 ACM/SIGDA ninth international symposium on Field programmable gate arrays*
92. Dhanachandra N, Manglem K, Chanu YJ (2015) Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput Sci* 54:764–771
93. Nilashi M et al (2021) What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis. *Telematics Inform* 64:101693
94. Nilashi M et al (2019) A hybrid method with TOPSIS and machine learning techniques for sustainable development of green hotels considering online reviews. *Sustainability* 11(21):6013
95. Zibarzani M et al (2022) Customer satisfaction with restaurants service quality during COVID-19 outbreak: a two-stage methodology. *Technol Soc* 70:101977
96. Nilashi M et al (2021) Recommendation agents and information sharing through social media for coronavirus outbreak. *Telematics Inform* 61:101597
97. Jeacle I, Carter C (2011) In TripAdvisor we trust: rankings, calculative regimes and abstract systems. *Acc Organ Soc* 36(4–5):293–309
98. Ma Y et al (2019) Operation flexibility evaluation and its application to optimal planning of bundled wind-thermal-storage generation system. *Electronics* 8(1):9
99. Flake GW, Lawrence S (2002) Efficient SVM regression training with SMO. *Mach Learn* 46(1):271–290
100. Nunno L (2014) Stock market price prediction using linear and polynomial regression models. Computer Science Department, University of New Mexico, Albuquerque, NM, USA
101. Nguyen H et al (2020) Predicting blast-induced ground vibration in open-pit mines using vibration sensors and support vector regression-based optimization algorithms. *Sensors* 20(1):132
102. Kashaninejad M, Dehghani A, Kashiri M (2009) Modeling of wheat soaking using two artificial neural networks (MLP and RBF). *J Food Eng* 91(4):602–607
103. Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140
104. Bhesdadiya R et al (2016) Training multi-layer perceptron in neural network using whale optimization algorithm. *Indian J Sci Technol* 9(19):28–36
105. Aljarah I, Faris H, Mirjalili S (2018) Optimizing connection weights in neural networks using the whale optimization algorithm. *Soft Comput* 22(1):1–15
106. Sharpley R (2009) Tourism and development challenges in the least developed countries: the case of The Gambia. *Curr Issue Tour* 12(4):337–358
107. Nguangchaiyapoom S, Yongvanit S, Sripun M (2012) Community-based tourism management of ban prasat, non sung district, nakhon ratchasima province. *Thailand Humanities and Soc Sci* 29(3):191–208
108. Formica S, Uysal M (2001) Segmentation of travelers based on environmental attitudes. *J Hosp Leis Mark* 9(3–4):35–49
109. Jang SC, Morrison AM, O'Leary JT (2002) Benefit segmentation of Japanese pleasure travelers to the USA and Canada: selecting target markets based on the profitability and risk of individual market segments. *Tourism Manag* 23(4):367–378
110. Prayag G et al (2015) Segmenting markets by bagged clustering: Young Chinese travelers to Western Europe. *J Travel Res* 54(2):234–250
111. Lee I, Shin YJ (2020) Machine learning for enterprises: Applications, algorithm selection, and challenges. *Bus Horiz* 63(2):157–170
112. Cheung K-W et al (2003) Mining customer product ratings for personalized marketing. *Decis Support Syst* 35(2):231–243
113. Martinez-Torres MdR, Toral S (2019) A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Manag* 75:393–403
114. Martín CA et al. (2018) Using deep learning to predict sentiments: case study in tourism. *Complexity*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.