Hindawi
Advances in Civil Engineering
Volume 2023, Article ID 6029121, 8 pages
https://doi.org/10.1155/2023/6029121

*Research Article*

# Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm

**Leila Taherkhani,[1] Amir Daneshvar ,[2] Hossein Amoozad Khalili ,[3] and Mohamad Reza Sanaei[4]**

[1]*Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[2]*Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[3]*Department of Industrial Engineering, Sari Branch, Islamic Azad University, Sari, Iran*
[4]*Department of Information and Technology Management, College of Management and Economics, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

Correspondence should be addressed to Amir Daneshvar; a_daneshvar@iauec.ac.ir

Received 12 December 2022; Revised 17 June 2023; Accepted 10 August 2023; Published 24 August 2023

Academic Editor: Afshar A. Yousefi

The ability of hotels to differentiate themselves from competitors and continue to operate profitably depends on their ability to retain their customers by building long-term and permanent customer relationships. Technological developments in recent years have made it possible for companies to predict their customers' behavior by accessing their opinions faster and preventing them from churning. Managing customer churn prediction projects has become an important issue today, especially in the hotel industry. Therefore, this research seeks to analyze projects that predict the churn of hotel customers to provide a model to help hotel managers in this field. In this research, an approach based on text mining on customers' comments in the Persian language is presented, which uses the random forest algorithm for classification that was considered the most effective method to solve this problem. In this model, to increase the efficiency of the proposed method in compare with existing works, the gravitational search algorithm was used to select the useful features, and the differential evolution algorithm was used to adjust the parameters of the classification method. The dataset of this research is the collected data from the customer database on social networks and hotels' websites, especially the hotels on Kish Island in Iran. The results of this research showed that after the implementation of the preprocessing operations, the method of adjusting the parameters and removing the unimportant features, the model's accuracy increased significantly. The precision, recall, F1, and accuracy criteria were 0.77, 0.76, 0.76, and 0.77, respectively.

## 1. Introduction

Customer churn prediction (CCP) is one of the most critical problems for a healthy growing business, regardless of size. CCP allows professionals to estimate the number of customers who abandon a company's product or service in a given period of time and take the necessary actions to retain them. In various markets, customers can quickly terminate their subscriptions with suppliers and switch to competing organizations for increased service quality and lower prices [1, 2]. Accurately predicting customer churn can effectively help customer retention and economical marketing activities and, therefore, can lead to significant savings for suppliers. Extensive research has been done in this field in sectors such as telecommunications and the hotel industry is no exception.

Online customer feedback tools enable organizations to generate and share content and feedback through social media. According to the report SiteMinder [3], 96% of hotel guests consider users' comments essential when searching for a hotel, and 79% read between 6 and 12 users' comments before making a decision. Therefore, reviews of users' comments affect hotel guests' decision-making and hotels' performance [4], and the value of user-generated content should be further investigated. This issue provides various opportunities for both companies and users. For example, in the hotel industry, there is a growing interest in analyzing the opinions of different guests and finding hidden patterns or influencing factors [5, 6]. Therefore, it is imperative to check customer churn based on text mining in the hotel industry. At present, the literature on big data mainly focuses on techniques such as sentiment analysis, latent Dirichlet allocation, regression modeling, and others. Despite these valuable contributions, new methods to understand the hidden concepts in unstructured data for different hospitality fields are needed. While there have been many studies and research projects on big data and business analysis, few research studies have investigated the actual content produced by customers on social media and the factors affecting their experience and flow of customers. In addition, how to use analytics tools to analyze user-generated content has not been widely studied, so this research has been addressed how to combine techniques such as text mining and random forest algorithm for CCP in the hotel industry.

The structure of the article is as follows: in the second part, we examine the background of the research. In the third part, we present the proposed system and research methodology. In the fourth part, we describe the dataset and the tools used for data analysis. In the fifth section, we present the results of data analysis in the form of tables and graphs. In the sixth part, we give explanations about managerial insights and practical implications, and in the seventh and final part of the article, we will have conclusions and suggestions.

## 2. Literature Review

During the last decade, CCP has received increasing attention to survive in a competitive and global market [7]. Companies should strive for models that can accurately identify potential churn customers. This issue becomes more critical with the development of information technology. During the last decade, many experts and academics were interested in this topic and paid attention to it. Several methods presented in recent years will be reviewed below. The primary purpose of the study conducted by Dursun Cengizci [8] was to predict the customers' behavior of a hotel business using machine learning methods in its customer database. In this context, CCP was applied using the data of regular customers of a hotel chain in Antalya, which has three five-star hotels, and logistic regression and random forest algorithms were compared. According to the findings of this study, the random forest algorithm performs better. It could accurately predict 80% (area under the curve (AUC) 0.80) of repeat customers who were likely to leave within the next 3 years.

To investigate the causes of customer churn in Ctrip agency, Zhao et al. [9] used the current and potential value in the customer value system to determine the influencing factors. This research used the random forest algorithm to build the Ctrip CCP model, and the confusion matrix and receiver operating characteristic (ROC) curve were used to evaluate the model. The results showed that the random forest algorithm can better solve the classification problem of CCP, and the accuracy of the prediction model reached 94%. In the research of Han [10], the customer churn of hotel reservation websites has been investigated in China. In this research, logistic regression and random forest algorithms were used to identify the characteristics that affect customer churn. The experimental evaluation of this research showed that the model had an accuracy of 78.9%.

In the research of Yang [11], Ctrip hotel customers' data were analyzed. First, a supervised feature selection method was used to select features that had a significant impact on customer churn. Then the best model was chosen from logistic regression, support vector machine, decision tree, random forest, GBDT, XGBoost, and LightGBM. A subset of optimal models was selected, and then integration of the optimal models was performed. Experimental results showed that multimodel fusion has higher accuracy and stability.

The study done by Christodoulou et al. [12] solved the problem of revisiting tourists from the point of view of big data analysis. The applied method used topic modeling, word embedding, XGBoost, and random forest classification algorithms. The data were collected from TripAdvisor. Topics were generated using STM topic modeling and information retrieval using Word2vec. The learned model achieved satisfactory performance. The XGBoost classification model achieved a prediction accuracy of 84% and an AUC of 90% for the study of tourists revisiting two to three-star hotels and a prediction accuracy of 89% and an AUC of 90% for four to five-star hotels. The goal of research by Gartvall and Skånhagen [13] was to predict hotel cancellations using machine learning and analyze the factors that have the most significant impact. The data were provided by a hotel in the Gothenburg area. The machine learning algorithms used in the thesis were random forest, XGBoost, and logistic regression. The main findings of this research showed that random forest is the best-performing model in hotel data, with an accuracy of nearly 80%. In the study conducted by Oh et al. [14], deep learning techniques were combined with expectation-confirmation theory to predict customer satisfaction with hotel services. The results showed that this model achieved an accuracy of 83.54%.

In the article by Nagaraju and Vijaya [15], a method was developed to identify the prediction of customer churn in the insurance sector using meta-heuristics and bagging and boosting techniques. In this research, an approach based on meta-heuristic feature selection was proposed to identify effective features. In this study, the combination of firefly enhanced with boosting group classifiers achieved the highest accuracy of 97.12. In the research conducted by Lalwani et al. [16], the prediction of customer churn in the telecommunications industry was made by using machine learning techniques. After data preprocessing, feature selection was made using the gravity search algorithm. In the prediction process, the most common models, including logistic regression, naive Bayes, support vector machine, random forest, and decision trees, were used. K-fold cross-validation was used to adjust hyperparameters and prevent overfitting. The results showed that AdaBoost and XGBoost classifiers have the highest accuracy of 81.71% and 80.8%, respectively. In the research done by Wu et al. [17], a tree-based mechanism was presented that considers temporal and behavioral information separately. Extensive tests in this research

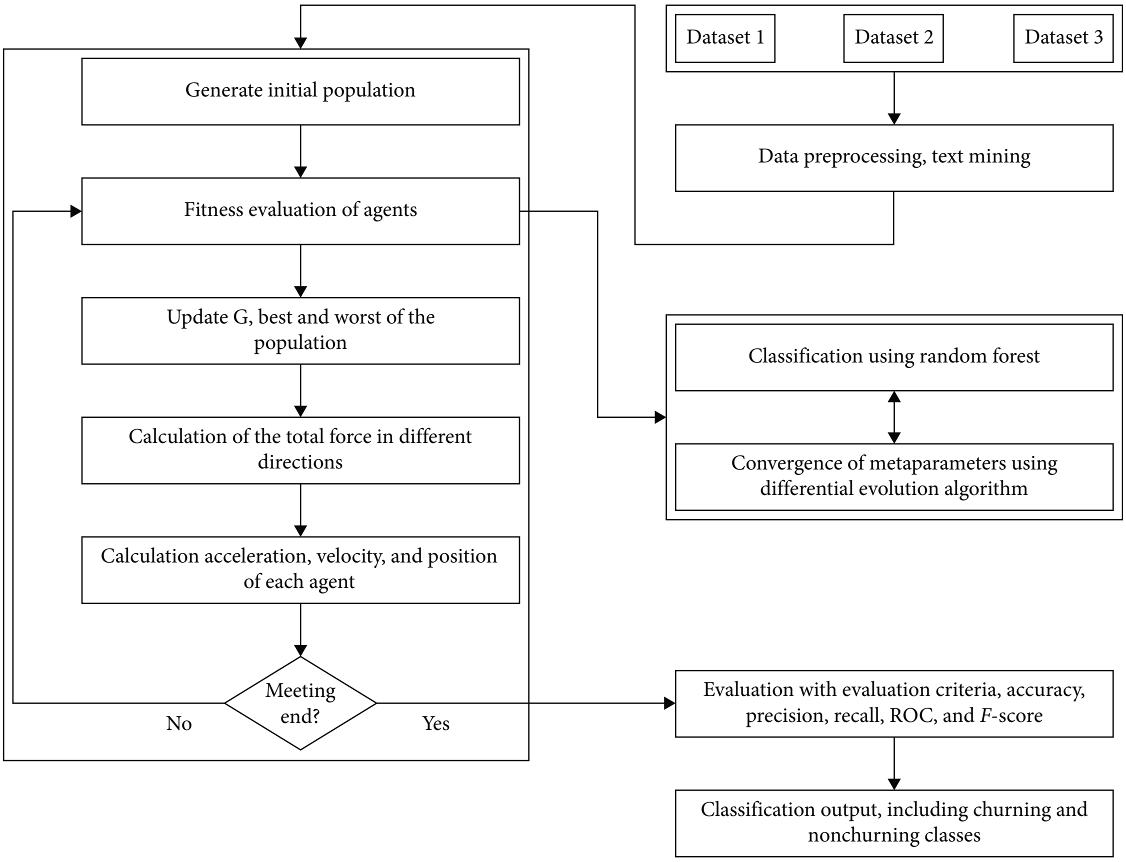
Dataset 1
Dataset 2
Dataset 3
Data preprocessing, text mining
Generate initial population
Fitness evaluation of agents
Update G, best and worst of the population
Calculation of the total force in different directions
Calculation acceleration, velocity, and position of each agent
Meeting end?
No
Yes
Classification using random forest
Convergence of metaparameters using differential evolution algorithm
Evaluation with evaluation criteria, accuracy, precision, recall, ROC, and *F*-score
Classification output, including churning and nonchurning classes

Figure 1: Proposed system.

showed that the proposed system achieved an *F* score of 82.72 and an AUC of 93.75.

Based on the background of the research, the previous studies in this field were mainly on the English language, and less attention was paid to other languages. Therefore, in this research, Persian data available on online platforms have been used, so the purpose of this research is to address this gap and predict customer churn using analytics of user-generated online Persian content. Also, this research intends to analyze the text content produced by customers through combining text mining methods and random forest algorithms and develop a model for CCP in the hotel industry.

The merits of the proposed method have listed as follows:

(i) We have applied the combined techniques, gravitational search algorithm to perform feature selection, and differential evolution algorithm to adjust the parameters of the classification method. Then we have compared those with the proposed method.

(ii) As the considered dataset is includes users' online comments in the Persian language, we have applied text mining techniques for preprocessing of data, such as case folding, tokenization, stop-words, stemming, and term frequency—inverse document frequency (TF-IDF).

(iii) We have applied some of the famous machine learning techniques for classification, which are used for predictions like gradient boosting classifier, naïve Bayes, decision tree, and KNN. Finally, we have compared those with the proposed classification method.

(iv) Then we have evaluated the algorithms using the confusion matrix and ROC, which have been mentioned in the form of figures and tables, in order to compare which algorithm performs best for this purpose.

Authors' contribution is using the gravitational search algorithm to select the useful features and combining it with the differential evolution algorithm to adjust the parameters of the classification method. The random forest algorithm was considered to this purpose. Then we evaluated the results through the confusion matrix and ROC, which have been mentioned in the form of figures and tables in order to compare which algorithm performs best for this purpose.

## 3. Proposed System

The proposed system in this research uses a combination of random forest, gravitational search algorithms, and differential evolution algorithms to predict customer churn. The diagram of the proposed system is shown in Figure 1. The details of this method are mentioned below. The reason for choosing the random forest method for learning the proposed system is the proper performance of this method compared to other machine learning methods in solving the research problem. The proof is provided in Section 5.

Table 1: Pseudo code of the differential evolution algorithm.

```
Pseudo code of differential evolution algorithm
Initial population with randomly generated individuals
Fitness evaluation of all individuals in the population
While the final conditions are not met,
  Create the mutant vectors using the mutation strategy
  Create trial vectors by combining noisy vectors with parent vectors
  Evaluate trial vectors with their fitness values
  Select winning vectors as individuals in a new generation
End while
```

As shown in Figure 1, first, the process of text mining and data preprocessing is done. In this section, operations like case folding, tokenization, stop-words, stemming, and TF-IDF are performed on the primary text data. Each of them is described below. To increase the efficiency of the proposed system, the selection of effective features in this research is made by the gravitational search algorithm. The main reason for using the feature selection method is to identify essential variables to reduce data dimensions and increase classification accuracy. A large number of predictive variables leads to a decrease in model accuracy. In this research, by implementing the gravity search algorithm, it is determined which features are valuable for solving the problem of predicting customer churn in the hotel industry.

Another important issue in obtaining proper performance from classification techniques is the correct setting of their parameters. The importance of the parameters of each learning model and problem-solving method, especially artificial intelligence methods that have been created to simplify problem-solving, is inevitable. The optimal values of these parameters, which generally depend on the characteristics of the problem, have a significant impact on the performance of the mentioned methods and a better search of the solution space. The random forest method has been increasingly used in various sciences and has been more successful than other existing methods in many fields. However, this method, like other learning models, is known to be sensitive to its parameters, and determining the appropriate combination of parameters has a significant impact on the final implementation of the algorithm and the results. Considering the challenge of determining the best values of the parameters, in this research, the differential evolution algorithm is also used to adjust the parameters of the random forest.

*3.1. Data Preprocessing.* The first section of the proposed method is text preprocessing. In this order, the texts are tokenized first; then, they enter the stage of removing stop words. At this stage, words that are repeated a lot in every document and do not have any meaning will be removed from the document. In the Case Folding step, all words are considered the same in terms of uppercase or lowercase letters. This step is done so that if a word is repeated several times with the same form but with different uppercase and lowercase letters, it is considered once in the modeling [18].

The next stage in the preprocessing step is stemming. The purpose of this stage is to harmonize the form of the words in the documents. With the help of stemming methods, words that are similar in terms of concept and differ from each other only in appearance are placed in one group and are considered features [19]. In the next stage, the goal is to find the appropriate weight of each word according to the TF-IDF weighting methods. The TF-IDF method is a standard weighting method in text mining studies that assigns the right weight to a word based on the number of repetitions of that word in each document and the number of repetitions in all documents and is calculated from Equation (1). In this regard, $t_k$ refers to the $k$th word, $d_i$ refers to the ith document, $N$ refers to the total number of existing documents, and $d_k$ refers to the number of documents with the term $t_k$ [20].

$$\text{TF-IDF}(t_k, d_i) = \text{TF}(t_k, d_i) \times \log\left(\frac{N}{d_k}\right). \tag{1}$$

*3.2. Classification Model, Random Forest.* The random forest method was presented by Breiman [21] as a new development method for decision trees. The general principles of ensemble training techniques are based on the assumption that their accuracy is higher than that of other singular training algorithms. Because it is a combination of several prediction models. It is more accurate than a single model and reduces existing weaknesses [21]. Several decision trees are used in this algorithm. A subset of data is given to each tree. These trees can make decisions and build their classification model with this subset of data [22]. The random forest algorithm is currently one of the best learning algorithms, and due to its good performance in solving the problem of customer churn, it was chosen for classification in this research.

*3.3. Differential Evolution Algorithm.* The differential evolution algorithm is presented to overcome the primary defect of the genetic algorithm, i.e., the lack of local search. The main difference between the genetic algorithm and the differential evolution algorithm is in the order of the mutation and recombination operators, as well as how the selection operator works in this algorithm. This algorithm uses a differential operator to generate new answers, which causes the exchange of information between members of the population. One of the advantages of this algorithm is having a memory that keeps the information about suitable answers in the current population. The pseudo-code of the differential evolution algorithm is presented in Table 1 [23].

*3.4. Gravitational Search Algorithm.* The gravitational search algorithm is a population-based and iteration-based stochastic

Table 2: Pseudo code of gravitational search algorithm.

Pseudo code of the gravitational search algorithm

```
Search space identification, initialization of parameters
Random initialization of agents.
While the final conditions are not met, do the following steps:
    Fitness evaluation of agents
    Update G($t$), best($t$), worst($t$) and Mi($t$) for $i = 1,2, \ldots, N$.
    Calculation of the total force in different directions
    Calculation of acceleration and velocity
    Update the position of each agent.
    Returning the best solution found.
End While
```

Table 3: Comparison of the impact of feature selection and parameter adjustment mechanism in the proposed system.

| Methods | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| Random forest + gravitational search algorithm for feature selection | 0.75 | 0.75 | 0.76 | 0.76 |
| Random forest + differential evolution algorithm for parameters adjustment | 0.74 | 0.75 | 0.76 | 0.74 |
| Proposed method | 0.77 | 0.76 | 0.76 | 0.77 |

meta-heuristic algorithm. This algorithm is inspired by nature to solve continuous optimization problems. The main idea of this algorithm is to simulate Newton's law of gravity and laws of motion on a population of masses in a constant $n$-dimensional space. In this algorithm, agents are considered objects, and their performance depends on their mass. Objects attract each other with the force of gravity, which causes the general movement of all objects toward objects with a heavier mass. Heavy masses represent better solutions and move slower than lighter masses. The pseudo-code of the gravitational search algorithm is presented in Table 2 [24].

## 4. Tools and Datasets

The information for this research was obtained from the customers' databases on social networks and hotels' websites, especially the hotels on Kish Island, from five-star to three-star hotels, and was collected in the form of textual content produced by customers. This dataset has more than 6,000 records and is in the Persian language. Among themes, 2,374 cases belong to customers who may churn, and 3,626 cases are related to nonchurning customers. Considering that the collection of users' comments is the first step of the desired implementation in this research, the comments of customers of Persian websites are first extracted with the help of the Python programing language. In the following, the facilities provided by the Python language were used for preprocessing, preparation of comments, and modeling.

## 5. Results

The results of this study are examined in this section. As already mentioned, the information for this research was obtained from the customers' databases on social networks and the websites of Kish Island hotels from five stars to three stars. It was collected in the form of text comments produced by customers. This dataset has more than 6,000 records in the Persian language. Among them, 2,374 cases belong to customers who may churn, and 3,626 cases are related to non-churning customers. In the first part of the evaluation of the results, an experiment was designed to check the effectiveness of the feature selection mechanisms and the parameter adjustment mechanisms. In this section, in the first case, the results of the proposed system are evaluated only with the feature selection mechanism. The best possible parameters are selected by the gravitational search algorithm. In addition, in this mode, the parameters of the random forest are chosen manually. In the second case, the results of the proposed system are evaluated only by the parameter adjustment mechanism. So, in this section, all the features of the dataset are used to solve the problem, and the differential evolution algorithm is used to adjust the parameters of the random forest.

The results of each mode are shown in Table 3 and were compared in terms of accuracy, recall, F1, and precision. For a better comparison, in Figure 2, the results were shown with the help of graphs. According to the results of this test, the feature selection mechanism has a more significant effect on improving the performance of the proposed system.

In the next experiment, in addition to the proposed system, several basic machine learning methods were applied to the dataset to check the proposed system's performance. Therefore, the methods of random forest, gradient boosting classifier, naive Bayes, decision tree, and KNN were applied to solve the customer churn problem in the hotel industry. The results from each of the methods are in Table 4. Methods were compared, according precision, recall, F1, and accuracy criteria. Figures 3–6 compare the results obtained from the mentioned methods in different ways.

As shown in these figures, the proposed system of this research has performed better than other methods in the precision, F1, and accuracy criteria with a big difference.

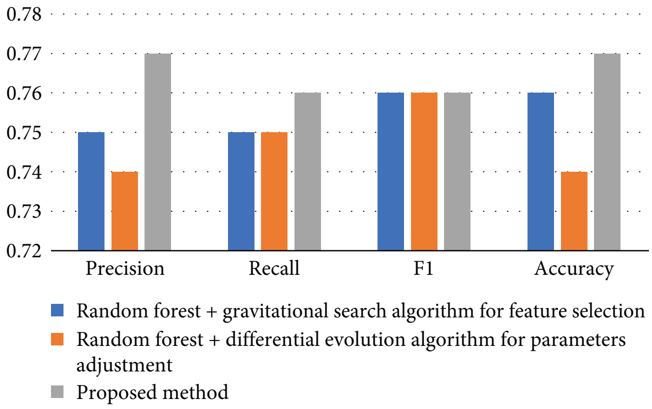

Figure 2: Comparison of the impact of feature selection and parameter adjustment mechanism in the proposed system.

Table 4: Performance comparison of random forest, gradient boosting classifier, naive Bayes, decision tree, KNN methods, and the proposed system.

| Methods | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| Random forest | 0.73 | 0.78 | 0.75 | 0.74 |
| Gradient boosting classifier | 0.74 | 0.74 | 0.74 | 0.74 |
| Naive Bayes | 0.76 | 0.71 | 0.73 | 0.74 |
| Decision tree | 0.68 | 0.66 | 0.67 | 0.68 |
| KNN | 0.69 | 0.65 | 0.67 | 0.68 |
| Proposed method | 0.77 | 0.76 | 0.76 | 0.77 |

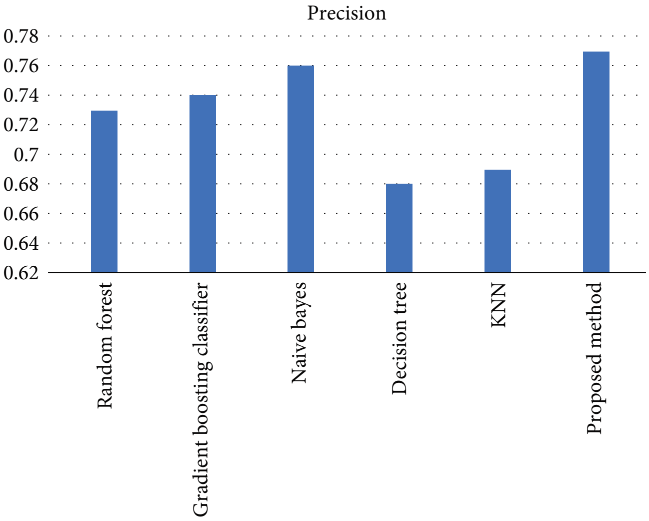

Figure 3: Comparison of the precision of random forest, gradient boosting classifier, naive Bayes, decision tree, KNN methods, and the proposed system.

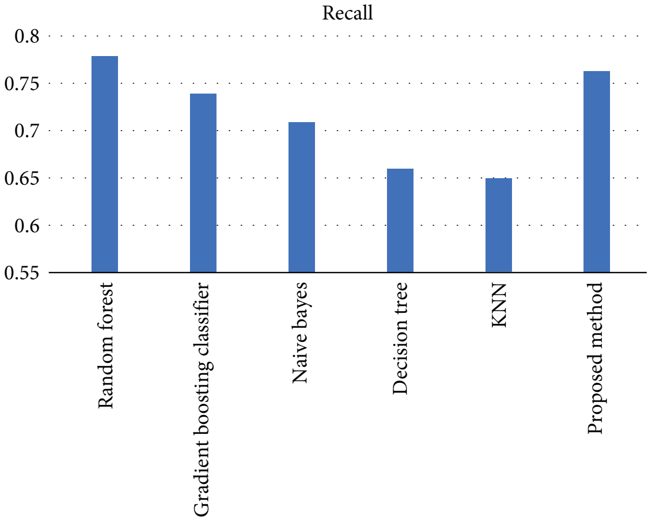

Figure 4: Comparison of recall of random forest, gradient boosting classifier, naive Bayes, decision tree, KNN methods, and the proposed system.

The primary random forest method has been better than the proposed method only in the recall criterion. However, this difference is slight, and due to the superiority of the proposed method in the other three criteria, it can be ignored. After the proposed system, among the basic methods, the random forest method had the best performance, and this issue itself indicates the appropriate choice of classification method in the proposed system. The primary random forest method has performed better than other basic methods regarding recall, F1, and accuracy. Only the naive Bayes method has

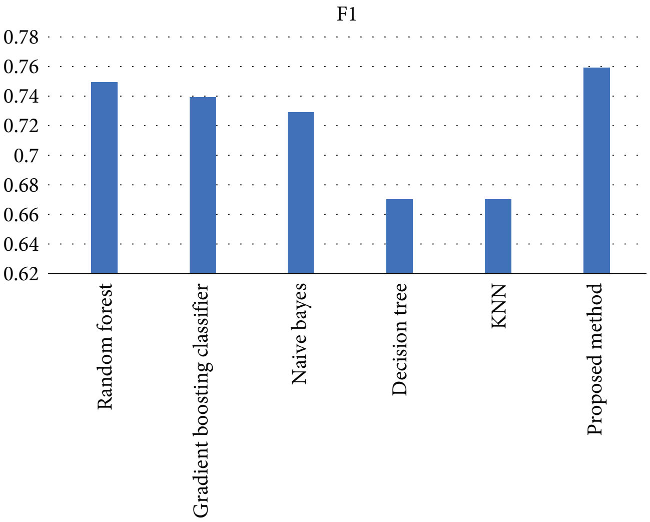

Figure 5: Comparison of F1 of random forest, gradient boosting classifier, naive Bayes, decision tree, KNN methods, and the proposed system.

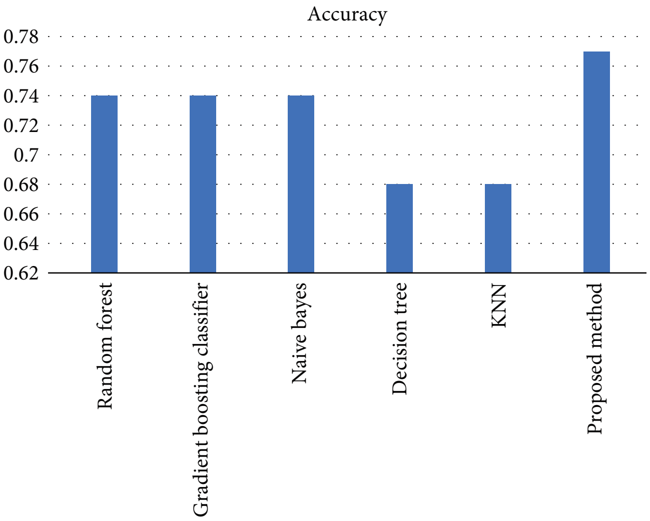

Figure 6: Comparison of the accuracy of random forest, gradient boosting classifier, naive Bayes, decision tree, KNN methods, and the proposed system.

served better than the primary random forest method in the precision criterion.

## 6. Managerial Insights and Practical Implications

By using the method presented in this research, it is possible to achieve reasonable accuracy in predicting the loss of customers. Therefore, the managers of the hotel industry, using the model presented in this research, should identify customer churn early and plan to solve the existing problems. By applying the model introduced in this research to the Persian dataset, it is possible to find the factors influencing the churn of customers. By having these components, the management of the relevant industry can focus on improving the services that lead to customer churn.

The solution proposed in this research can be applied to all kinds of Persian texts for different industries. Using this method and according to the optimizations, we can expect to achieve the appropriate accuracy for predicting customer churn.

## 7. Conclusions and Outlook

In this research, a hybrid approach based on text mining, random forest, the gravity search algorithm, and differential evolution is presented to solve the problem of predicting customer churn in the hotel industry. The proposed model helps to extract and review the key performance indicators from the massive volume of comments from hotel customers so that the decision-making process can be done more effectively. The following are the findings of the research:

(1) The results of the research and comparisons showed that the proposed system of this research has a good performance compared to the compared methods and has reached an accuracy difference of 0.03.

(2) The proposed method was applied to the set of opinions of hotel customers on Kish Island and the precision, recall, F1, and accuracy criteria; the results were 0.77, 0.76, 0.76, and 0.77, respectively.

(3) One of the reasons for the superiority of the proposed system is the selection of the random forest method, which, according to the obtained results, has performed better than other basic methods to solve this problem. One of the reasons for the high accuracy of the proposed model compared to the other methods is the feature selection method using the gravity search algorithm. The proposed system's importance of each extracted feature is checked to solve the problem. This method leads to better training of the model by selecting a subset of essential features, thus increasing the model's accuracy.

(4) The random forest parameter adjustment mechanism using the differential evolution method improves the performance of the random forest method.

(5) According to the tests, the feature selection process has a more significant impact on system efficiency than the parameter setting and is considered an essential part of problem-solving.

This research, like other research, has limitations. Part of the limitations of the present research are related to the collected dataset, among which the lack of Persian text data in this field and the exclusiveness of the data to several specific hotels can be mentioned.

As a suggestion for the future, the proposed model can be implemented in a cloud-based environment or the Internet of Things to predict travel planning and so on. In addition, ensemble methods can increase the final accuracy of the system by combining the advantages of machine learning techniques.

## Data Availability

Data supporting this research article are available from the corresponding author or first author upon reasonable request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] Y. Li, B. Hou, Y. Wu, D. Zhao, A. Xie, and P. Zou, "Giant fight: customer churn prediction in traditional broadcast industry," *Journal of Business Research*, vol. 131, pp. 630–639, 2021.

[2] J. Uthayakumar, N. Metawa, K. Shankar, and S. K. Lakshmanaprabu, "Financial crisis prediction model using ant colony optimization," *International Journal of Information Management*, vol. 50, pp. 538–556, 2020.

[3] SiteMinder, "How to influence travelers: why reviews are the golden egg at your hotel," 2017, https://www.siteminder.com/r/marketing/hotel-online-reviews/influence-travellers-reviews-hotel.

[4] P. Bligh and D. Turk, *CRM Unplugged: Releasing CRM's Strategic Value*, Wiley, Hoboken, NJ, USA, 2004.

[5] M. Lee, M. Jeong, and J. Lee, "Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: a text mining approach," *International Journal of Contemporary Hospitality Management*, vol. 29, no. 2, pp. 762–783, 2017.

[6] Z. Xiang, Z. Schwartz, J. H. Gerdes Jr., and M. Uysal, "What can big data and text analytics tell us about hotel guest experience and satisfaction?" *International Journal of Hospitality Management*, vol. 44, pp. 120–130, 2015.

[7] N. Gordini, "Market-driven management: a critical literature review," *Symphonya Emerging Issues in Management*, no. 2, pp. 95–107, 2010.

[8] A. Dursun Cengizci, "Otel işletmelerinde kayıp müşteri tahminlemesi," 2020, https://acikbilim.yok.gov.tr/handle/20.500.12812/32494.

[9] Z. Zhao, W. Zhou, Z. Qiu, A. Li, and J. Wang, "Research on ctrip customer churn prediction model based on random forest," in *Business Intelligenceand Information Technology. BIIT 2021, vol 107 of Lecture Notes on Data Engineering andCommunications Technologies*, A. E. Hassanien, Y. Xu, Z. Zhao, S. Mohammed, and Z. Fan, Eds., pp. 511–523, Springer, Cham, 2021.

[10] S. Han, "A study on a predictive model of customer defection in a hotel reservation website," in *MATEC Web of Conferences*, vol. 228, p. 1009, EDP Sciences, 2018.

[11] W. Yang, "Research on early warning of customer churn based on mutual information and integrated learning—taking ctrip as an example," *Academic Journal of Computing & Information Science*, vol. 5, no. 3, pp. 785–800, 2022.

[12] E. Christodoulou, A. Gregoriades, M. Pampaka et al., "Application of classification and Word Embedding Techniques to Evaluate Tourists' Hotel-revisit Intention," in *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pp. 216–223, SciTePress, 2021.

[13] E. Gartvall and O. Skånhagen, "Predicting hotel cancellations using machine learning," 2022, https://gupea.ub.gu.se/bitstream/handle/2077/70742/gupea_2077_70742_1.pdf?sequence=1.

[14] S. Oh, H. Ji, J. Kim, E. Park, and A. P. del Pobil, "Deep learning model based on expectation-confirmation theory to predict customer satisfaction in hospitality service," *Information Technology & Tourism*, vol. 24, no. 1, pp. 109–126, 2022.

[15] J. Nagaraju and J. Vijaya, "Boost customer churn prediction in the insurance industry using meta-heuristic models," *International Journal of Information Technology*, vol. 14, pp. 2619–2631, 2022.

[16] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, pp. 271–294, 2022.

[17] X. Wu, P. Li, M. Zhao, Y. Liu, R. G. Crespo, and E. Herrera-Viedma, "Customer churn prediction for web browsers," *Expert Systems with Applications*, vol. 209, pp. 118–177, 2022.

[18] C. C. Aggarwal and C. Zhai, *Mining Text Data*, Springer, Boston, MA, USA, 2012.

[19] C. Ramasubramanian and R. Ramya, "Effective pre-processing activities in text mining using improved porter's stemming algorithm," *International Journal of Advanced Research in Computer and Communication*, vol. 2, no. 12, pp. 4536–4538, 2013.

[20] S. M. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining*, Springer, London, 2010.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[22] K. Senagi, N. Jouandeau, and P. Kamoni, "Using parallel Random Forest classifier in predicting land suitability for crop production," *Journal of Agricultural Informatics*, vol. 8, no. 3, pp. 23–32, 2017.

[23] M. Leon and N. Xiong, "Investigation of mutation strategies in differential evolution for solving global optimization problems," in *International conference on artificial intelligence and soft computing*, pp. 372–383, Springer, Cham, 2014.

[24] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "GSA: a gravitational search algorithm," *Information Sciences*, vol. 179, no. 13, pp. 2232–2248, 2009.