



OPINION MINING FOR HOTEL RATING THROUGH REVIEWS USING DECISION TREE CLASSIFICATION METHOD

Shelly Gupta

Assistant Professor

CSE dept., IPEC Ghaziabad, India

Shubhangi Jain

Student

CSE dept., IPEC Ghaziabad, India

Shivani Gupta

Student

CSE dept., IPEC Ghaziabad, India

Shruti

Student

CSE dept., IPEC Ghaziabad, India

Aakriti Chauhan

Student

CSE dept., IPEC Ghaziabad, India

Abstract: Websites like TripAdvisor has become an intricate part of our lives. It gives us various reviews about the hotels and helps us to decide which hotels to stay in or visit. Customer reviews play a major role in influencing others. Hence, these reviews become very important in controlling the quality of the hotel. The goal of this research is to collect all such reviews from the web and generate rating based on it with the help of a data mining classification method that is decision tree. The C4.5 decision tree method is applied for the above purpose using Tanagra Machine learning tool to generate the data model. The advantage of using Decision tree algorithm is that the rule set can be easily generated and by analyzing each level of the tree we can improve a particular service quality.

Keywords: Opinion Mining, Feature Extraction, Decision tree, Data modeling, Classification

1. INTRODUCTION

There has been a visible growth in the number of people who are dependent on the internet to find hotel rooms. They prefer this method because it is much convenient way than the traditional one. The hotels try to get the reviews of their customers by different means such as by using guest books, by taking surveys etc. but these efforts don't give valuable results as the customers are reluctant to give negative reviews to hotel authorities and complain about the dissatisfactory services.

The major challenge while running a hotel is to satisfy the customers. The reviews given by the customers play a major role on the success of a hotel. The customers tend to judge hotels on basis of a variety of categories. But some of these categories are common to both, the positive reviews as well as the negative reviews. With the vision of expansion organizations have started a better method to know about the customers' point of view by using online feedback review system. The reviews on internet are mostly unbiased and true. So now, the hotel management reaches out to these websites, blogs and portals to see where their hotel/property stands in such a competitive market. Therefore the objective of the research is to provide a strong mechanism of hotel searches and generate an optimum rating by extracting some elements from the reviews which help the customer and hotel management in the hotel evaluation.

However, the reviews on the Internet are in abundance. It is difficult for the management of any hotel to manually analyze/study each review and see if the customer is satisfied or not. Also the online availability of reviews allows other

customers to check which hotel provides best value for money and satisfaction.

But again, the hotel management analyzing the reviews needs some kind of summary of all the reviews in one place because it's not possible for them to read each and every review. Taking into consideration the above problem of handling such a huge amount of data, in this research paper opinion mining of the reviews written by customers is performed using data mining classification method. Various features are extracted from the reviews using opinion mining. The sentences from the reviews are classified into various aspects and then sentiment analysis is applied to them. This research thus helps any customer to see whether a particular property/hotel has satisfactory services and can influence the choice of customers in a positive way for their maximum satisfaction. It also helps the hotel management to study various aspects at which they lag and improve their services. Hotel management will get to know the areas which will lead to positive or negative rating. This would help them improve the negative aspects and to provide better quality management.

This paper is organized in the following manner- section (ii) includes the literature survey, section (iii) describes the proposed methodology, section (iv) reflects the experimental results and finally section (v) in which conclusion is explained.

2. LITERATURE REVIEW

This section provides the brief description of various research papers studied for this study. The given below table 1 represents the summarization of various methods applied for opinion mining.

TABLE I: METHODS APPLIED FOR OPINION MINING

S.No.	Title	Method	Description
1	Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews	Text mining	Katrina Barezina et. al. [1] used text mining techniques to examine what are the different factors that satisfy or dissatisfy a customer. A conclusion was drawn that if quality service was not delivered then the customer is definitely dissatisfied.
2	A novel approach to rate and summarize online reviews according to user-specified aspects	NGD, Clustering, delphi	Hsiao-Wei Hu et. al. [2] found that websites only provide customer reviews of products or services so that the user can search for information from online reviews that are predefined by the web site. Sometimes users may be interested in some aspects of a product or service that has not been predefined by the websites, it becomes very difficult to extract it. In their paper they have tries to solve this problem by using clustering.
3.	Sentiment Analysis for Hotel Reviews	Statistical polarity classification	Walter Kasper et. al. [3] developed such a system that was capable of retrieving user comments. The web based mining system helped the hotel staff to manage and monitor what was written about their houses.
4.	Opinion Mining of Online Customer Reviews	Apriori Algorithm	Patlammagari Gowtamreddy [4] has applied different techniques and algorithms to perform opinion mining and sentimental analysis to obtain effective information from customer comments.
5.	The Analysis and Prediction of Customer Review Rating Using Opinion Mining	Decision tree, naïve bayes	Wararat Songpan [5] proposed a system which help in the analysis of the open reviews and generated a rating using classifier model. The problem with comments is that they may be different from the rating, that results in biasing.
6.	A study of opinion mining and visualisation of hotel reviews	Opinion mining	Thomas H.Burnett et. al. [6] proposed a prototype for assessing customer opinion about the hotels. Evaluated why there might be a change in opinion and which are better geographical areas.
7.	Online Traveler Ratings: Impact on Determining the Top 25 Hotels	Sampling and instrumentation	Yaitza Enid Matos-Rodríguez [7] concluded that the obtained results could be used for hotel management. The manager can understand the importance of ratings and reviews and also analyzed the relationship between the ratings and the top 25 hotels of united states of America.
8.	The future of hotel rating	Studied current trends and their impact	Wouter Hensens [8] has proposed a system where relationship between online reviews and conventional rating system is studied and also given important points how the present and future trends will impact the customers and how they can select a particular hotel.
9.	The Impact of Users' "Online Reviews" and "Ratings" on Consumers' Behaviour toward Hotel Selection Factors	Qualitative and quantitative research methods	Shahab Pourfakhimi [9] aimed to identify what influences a consumer to select a particular hotel and how online reviews can affect it.
10.	A review of studies on luxury hotels over the past two decades	Population and sampling	Yin Chu, [10] aimed to study present research trends and to see how the study of luxury hotels are changing. They concluded that qualitative methods should be used rather than quantitative methods so as to understand the hotel industry in an extensive manner.

11.	The Factors Affecting Writing Reviews in Hotel Websites The Factors Affecting Writing Reviews in Hotel Websites	Empirical analysis	Asunur Cezar et. al. [11] have tried to find out what motivates the customers to write reviews. They analyzed that if the rating is high and the price is on the lower side, then the customer shows tendency to write reviews.
12.	Understanding the impact of online reviews on hotel performance	Empirical analysis	Paul Phillips et. al. [12] analyzed the data on some predefined attributes and concluded that positive reviews increased the demand of that hotel.

3. METHODOLOGY

For this research work the KDD process [13] is taken as the methodology. The various steps involved in KDD process are discussed below and shown in figure 1.

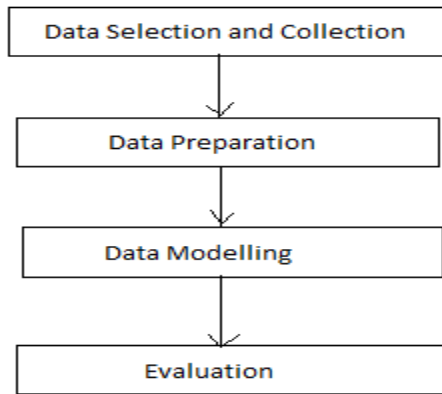


Figure1: Kdd Process

A. Data Selection and Collection

To achieve the desired objective of study the customers’ review data about the hotels which they have visited is selected and the data set was collected from the Trip Advisor website. The data collected was in the form of text comments. The data also included the overall rating of the hotel ranging from 1-5. Out of the entire data, 143 tuples (comments) have been selected and 14 attributes or features have been extracted based on their frequency of occurrence in the comments. The detail of selected attributes is shown below in table 2.

Table 2: Feature Selection From Frequent Words

Positive words	Frequency	Negative Words	Frequency
Great	228	Small	113
Nice	168	Problem	30
Clean	107	Noisy	23
Helpful	70	Expensive	18
Best	62	Bad	17
Comfortable	59	Disappointed	39
Friendly	58	Crowded	20

B. Data Preparation

On the basis of data, 14 attributes have been selected for classification. Seven of them are positive and seven are negative attributes. The following figures 2 and 3 shows the attributes selected under positive and negative categories respectively.

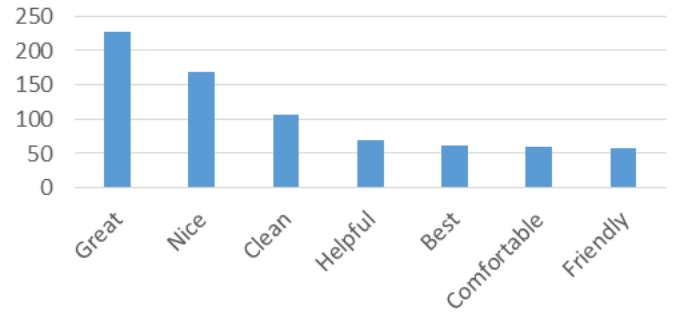


Figure 2: Words (Positive)

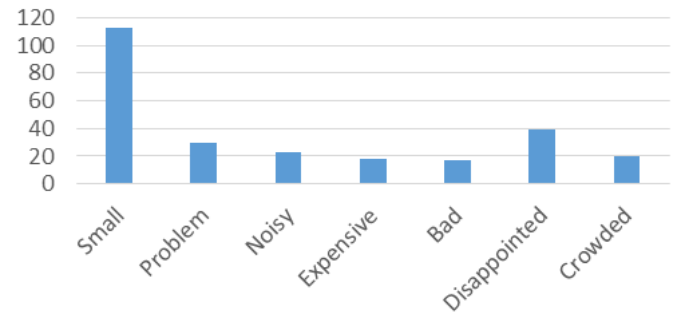


Figure 3: Words (Negative)

These attributes have been searched in the comments and their presence and absence in the comments have been denoted by true and false. If the overall rating is between 4-5 then the rating is taken to be positive else negative.

C. Model construction

The decision tree classification model has been used to classify the review into two classes- positive or negative. The decision tree C4.5 has been used which is easily represented in a tree form. The root node is selected and it further selects attributes as words. Every attribute already has a true or false value associated with it for each review.

The attribute “great” has been selected as the root node. The leaf nodes represent the class (positive/negative). Tanagra tool [14, 15] has been used to implement the decision tree C4.5 and has given 13 rule sets.

D. Evaluation

After data modeling, evaluation is done. The decision tree model is evaluated to derive conclusions. Rules are derived from the decision tree. From those rules, we try to find the attributes which create a major effect on the hotel rating and which are responsible for negative and positive rating. Accuracy of each rule is also evaluated to check their significance.

4. EXPERIMENTAL RESULTS

After applying the classification algorithm Decision tree C4.5 on the prepared dataset, we have obtained the following confusion matrix and tree through which some results have been derived.

TABLE 3: CONFUSION MATRIX

	Positive	Negative	Sum
Positive	59	20	79
Negative	15	20	64
Sum	73	70	143



Figure 4: Generated Decision Tree

On the basis of the generated decision tree and by the analysis of its patterns, following rules have been derived:

RULE 1: IF great = False, problem= False, small= False, Comfortable= False and nice= False THEN negative (Accuracy: 66.67%)

RULE 2: IF great = False, problem= False, small= False, Comfortable= False and nice= True THEN positive (Accuracy: 66.67%)

RULE 3: IF great = False, problem= False, small= False and Comfortable= True THEN positive (Accuracy: 100%)

RULE 4: IF great = False, problem= False and small= True THEN negative (Accuracy: 70.37%)

RULE 5: IF great= False and problem= True THEN negative (Accuracy: 90.91%)

RULE 6: IF great= true, disappointed=False, bad=False, helpful=False, problem= False, best=False and small= False, THEN positive (Accuracy: 73.08%)

RULE 7: IF great=True, Small=True and friendly=false THEN positive (Accuracy: 66.67%)

RULE 8: IF great=True, Small=True and friendly=true THEN negative (Accuracy: 60%)

RULE 9: IF great=true, Disappointed=False, Bad=False, Helpful=False, Problem=False and best=True THEN positive (Accuracy: 87.7%)

RULE 10: IF great=true, Disappointed=False, Bad=False, Helpful=False and Problem=True THEN positive (Accuracy: 80%)

RULE 11: IF great=true, Disappointed=False, Bad=False, Helpful=True THEN positive (Accuracy: 93.75%)

RULE 12: IF great=true, Disappointed=False and Bad=True THEN negative (Accuracy: 62.51%)

RULE 13: IF great=true and Disappointed=True THEN negative (Accuracy: 71.431%)

From the above rule set generated by using decision tree C4.5 classification model, we can conclude the following:

1. COMFORTABLE is such an attribute which when TRUE overrules the attribute GREAT. This can be seen by Rule 3 which has 100% accuracy.
2. GREAT is such an attribute such that when it is true, it overrules values almost all the other attributes to divert the review to positive. We can see this from rule 6,9,10 that give us accuracy>73%.
3. Even if GREAT is true, DISAPPOINTED is an attribute in a customer’s review that turns the review towards being negative. This is projected by rule 13 that gives accuracy>71.43%.
4. We can see in rule 11 that even if people are DISAPPOINTED by some services then also if the staff is HELPFUL, the review goes positive. The accuracy of this rule is 93.75%.

5. CONCLUSION

To improve services provided by a hotel, Opinion Mining is used. Here the model has been generated using Decision Tree algorithm. The accuracy of the decision tree is 76.22% which is remarkable to generate a system. The customers can read all the comments but it is time consuming. The model based on the above results can tell that the hotel has overall positive reviews or negative reviews by just picking some words from the whole review. Thus the generated model can help the customers and hotel management to take their decision for future. The accuracy of the system can be improve by using more efficient methods.

REFERENCES

- [1]. K.Berezina, A. Bilgihan, C. Cobanoglu, F. Okumus, "Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews", *Journal of Hospitality Marketing & Management*, Volume 25, Issue 1, 2016.
- [2]. HW. Hu, YL. Chen, PT. Hsu, "A Novel Approach to Rate and Summarize Online Reviews According to User-Specified Aspects", *Journal of Electronic Commerce Research*, Volume 17, Issue 2, 2016, Page 132.
- [3]. W. Kasper, M. Vela, "Sentiment Analysis for Hotel Reviews", *Proceedings of the Computational Linguistics-Applications Conference*, Poland, pp. 45-52, 2011.
- [4]. P. Gowtamreddy, "Opinion Mining of Online Customer Reviews", MTech thesis, National Institute of Technology, Rourkela, 2014.
- [5]. W. Songpan, "The Analysis and Prediction of Customer Review Rating Using Opinion Mining", *Software Engineering Research, Management and Applications (SERA)*, IEEE 15th International Conference, London, 2017.
- [6]. E. Bjorkelund, T H. Burnett, K. Norvag, "A Study of Opinion Mining and Visualization of Hotel Reviews", *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, Pages 229-238, Indonesia, 2012.
- [7]. Y. E. Matos-Rodríguez, "Online Traveler Ratings: Impact on Determining the Top 25 Hotels", MBA Student Scholarship, Johnson & Wales University, 2014.
- [8]. W. Hensens, "The Future of Hotel Rating, *Journal of Tourism Futures*", Volume 1, Issue 1, 2015.
- [9]. S. Pourfakhimi, "The Impact of Users' "Online Reviews" and "Ratings" on Consumers' Behaviour toward Hotel Selection Factors", *Information and Communication Technologies in Tourism*, pp. 104-114, 2014.
- [10]. Y. Chu, "A review of studies on luxury hotels over the past two decades", Unpublished master's thesis, Iowa State University, Iowa, United States, 2014
- [11]. H. Ögüta, A. Cezara, "The Factors Affecting Writing Reviews in Hotel Websites", *Procedia-Social and Behavioral Sciences*, Volume 58, Pages 980-986, 2012.
- [12]. P. Phillips, S. Barnes, K. Zigan, R. Schegg, "Understanding the Impact of Online Reviews on Hotel Performance - *An Empirical Analysis*", *Journal of Travel Research*, Volume 56, Issue 2, Pages 235-249, February 1, 2017.
- [13]. S. García, J. Luengo, F. Herrera, "Data Preprocessing in Data Mining", Springer, ISBN 978-3-319-10247-4.
- [14]. Tanagra - A free Data Mining Software, <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>.
- [15]. Tanagra Data Mining, https://www.youtube.com/watch?v=5MtsExnb_J0&t=36s