Opinion Mining and Sentimental Analysis of TripAdvisor.in for Hotel Reviews

Divyashree N[1], Santhosh Kumar K L[2], Jharna Majumdar[3]

[1]PG Student, Dept. of M.Tech CSE, Nitte Meenakshi Institute of Technology, Karnataka, India
[2]Assistant Professor, Dept. of M.Tech CSE, Nitte Meenakshi Institute of Technology, Karnataka, India
[3]Dean R&D, Professor & Head of the Department, Dept. of M.Tech CSE, Nitte Meenakshi Institute of Technology, Karnataka, India

---

Abstract - *There are several websites which are available on the website such as www.Tripadvisor.in which will permit the customers to contribute their view about hotels, tourist place, shopping website etc. The necessary and the information which is important about each hotel should be fetched from each hotel review and offered to the customer. In this paper, an online connection is established to the website www.Tripadvisor.in. The Hotels details are extracted with the required information. Data mining Techniques are used to analyze the positive and negative words from the reviews of all the Hotels on the TripAdvisor website.*

***Key Words*: Sentimental Analysis, Data mining Techniques, PAM Algorithm, J48 Algorithm, Opinion Mining**

1. INTRODUCTION

Data mining is the method of discovering unknown pattern and trend from a lot of databases. The data mining method is the procedure to extract modern information and employ it . The Data mining main point is concerning on the responsibility to find the useful and the data which is not identified in the dataset.

Opinion Mining (OM) or Sentimental Analysis (SA) is nothing but the study of what the people think or their opinion, attributes and the emotions towards the entity. The topics, events or individuals represent the entity. The entity is covered by reviews. Extracting and analyzes opinion of people about an entity is called Opinion Mining while the feeling or emotions can be expressed in the form of the text then it will analyzes which is known as the Sentimental Analysis.

The technology has been developed due to which, the huge number of customers will access website as a basis of knowledge. Several users do online communication for hotels, movies, shopping websites etc. There are several users share their data or knowledge about the movie, hotel or topic on website in the form of reviews. There are several websites which are available on the website such as www.Tripadvisor.in which will permit the customers to contribute their view about hotels, tourist place, shopping website etc. The huge amount of customers will express their taught or opinion towards the hotel in the form of reviews, there should require the mining of customer generate contents which are there in the reviews.

Many customers or users read the reviews which are written and use the review information for making the decision about the Movie, Restaurant, shopping website etc. It is very difficult for the users to read and realize all the review about each hotel on the website. The necessary and the information which is important about each hotel should fetch from each hotel review and offer to the customer.

In TripAdvisor website, there are many hotels. Many users will look for which are the best hotels in the website based on the reviews given to the hotel by the customers. Each time there is a need to read all the reviews of all hotels it will be time-consuming. Therefore, the essential and necessary information can fetch from the review given by the customers through Opinion Mining and Sentimental Analysis.

This paper presents the mining the opinion and analysis of sentimental analysis of hotel reviews. Section 2 contains Literature Survey, in Section 3 presents methodology, section 4 contains experiments performed and results obtained and at last conclusion and future work.

2. LITERATURE SURVEY

In [1], the mining of the Opinion can be classified based on users decision for the particular issue which is obtained from the reviews. Opinion mining is one which classifies the review wording as negative opinion or positive opinion. The summarization of opinion is the method of finding the aspects which are important for the topic and which are related sentences which are related to reviews due to which the summary can be characterized. There are three modules for the proposed system they are classification, summarization, and classification.

a. Hotel reviews are retrieved from a website like www.Tripadvisor.in by the technique of web crawling.
b. The positive review and negative review can be classified by means of SentiWordNet or machine learning classifiers algorithm.
   The pre-processed and sentence scores in the text of the classified review can be calculated in it.

In [2], the online reviews and evaluations are improved due to this it will be tricky for many customers will differentiate the necessary reviews which can be obtained from the not useful ones. The predictive and descriptive are the two categories that can be separated in the data mining technique. The statistical summarization is nothing but the descriptive mining of the data.

The opinion mining is having the Sentimental analysis as the technique in it. It will repeatedly collect and extract the sentiments in the sentences. Grammatically the sentences can be separated for extracting its support. Opinion expression should be calculated for the sentiment of the events to calculate the subjective scores. The presence of the opinion can be calculated by the Classification. There are many phases of the opinions such as grammatical errors, spelling etc.

In [3], the aim is to extract the necessary information in data mining from a large amount of the database. Clustering: It is one of the original systems in the data mining. Its main aim is to group the similar objects together. Were each group is called a cluster. Clustering principals are Homogeneity: elements are close to each other and Separation: elements are far to each other. Cluster types: Clustering types are center based, density based, computational clustering etc. Techniques of Clustering: They are classified into two types of the cluster and additional is a grid-based and density based method.

In [4], Data mining revolves roughly the job of judging useful, unidentified pattern in a data set. Clustering is a machine learning method in which the grouping of data points into a separate cluster. Two types of clustering algorithm are a Partitional and hierarchal cluster. Partitional categorize data point into the nonoverlapping cluster. In Partitional cluster each data point will assign to at least one and only 1 cluster. The K-means will compute the mean values and therefore this can apply only to knowledge sets that area unit are stringently numerical in nature. PAM is the replacement clustering algorithm which will work on categorical knowledge sets. PAM will work like to K-means, aside from changing and finding the cluster center.

In [5], Classification is a method to classify the data into the assured class base on several similarity or criteria. Classification is additionally called as supervised learning, because of the occurrence square measure gave with illustrious cluster label, a distinction to unsupervised learning during which the label square measure is unidentified.

J48 decision tree classifier*:* J48 is that the decision tree related algorithm and which is the extension of C4.5. By this technique the tree is constructed to create the categorization procedure in call tree the inside nodes of the tree will denote the check on the associate degree attribute, in which the ranch represents the results of the check, the top node is that the root and leaf nodes hold the category label.

In [6], The Data mining is one of the procedures to find out the motivating information, such as anomalies, patterns changes, significant structures and associations from a large amount of information stored in the database. Classification is one of the actions to construct the form of the module from the place of account that manages group label. The Algorithm for decision tree is used to discover the process that the attributes-vector behaves for the integer of the instance. J48 is the enlargement of ID3. The additional options of J48 should be continuous attribute worth ranges, accounting for missing ideals, beginning of rules, decision trees pruning etc. Classification algorithms are Instance Based for decision tree (J48), K-Nearest neighbor (IBK), Naive Bayes, Sequential Minimal Optimization (SMO) and Multi-Layer opinion which are compared by means of medium and organization correctness.

3. METHODOLOGY

In this paper, the online connection is established to the website www.Tripadvisor.in Online connection is established to the web page of TripAdvisor.in for Hotel Reviews using Jsoup. The link given is https://www.tripadvisor.in/Hotels-g297628-Bengaluru_Bangalore_District_Karnataka-Hotels.html. The extracted data of hotel web page is stored in the CSV file format. It contains six attributes and 109 instances. The six attributes are Hotel name, Reviews, Stars, Hotel Address, Hotel Link and Review Text and the 109 instances are the names of all the Hotels. Using the extracted data and the Data mining techniques such as Clustering and Classification will analyze and predict the positive and negative words. Partition around medoids [PAM] Algorithm is used for clustering the extracted data, and then it forms various clusters. The j48 algorithm is used for classification and it is applied to the clustered data. Then it analyses the positive and negative words from the reviews of all the Hotels on the TripAdvisor website. It will also display the top ten hotels on the TripAdvisor website.

3.1. Objective:

Instead of reading all the reviews from the TripAdvisor website for all the hotels finding all the positive and negative words will help the customers to decide how many positive and negative words are there for a particular hotel. The overview of Methodology and the proposed system flow is shown in the Fig -1

3.2. Extracting:

- *Input:* Online connection is established to website Ex.TripAdvisor.in to the given location Ex. Bangalore for Hotel Reviews. Initially, a new connection is created to the web page of TripAdvisor.in for Hotel Reviews using Jsoup.

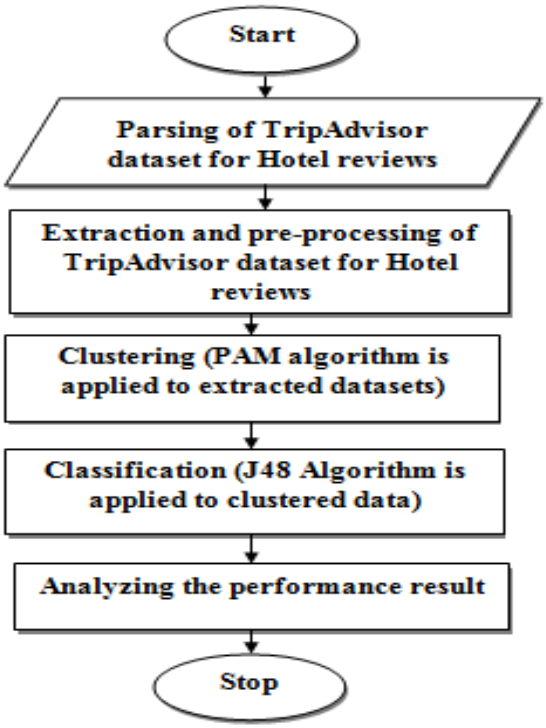

Fig -1: Methodology of Proposed System

After connecting, the HTML document is fetched, and then it is parsed and then finds data within it (screen scraping). The class name is specified to get the required HTML Document. The extracted data are hotel name, number of reviews, stars, hotel link, hotel address and reviews of the hotel.

- *Output:* Extracted data is obtained and stored in CSV file format.

3.3. Clustering:

Clustering is the technique of Data mining. The goal of the clustering is to split all the data elements to form similar objects to one group, each group is called cluster, in which all the items are similar to each other and dissimilar objects to the new group. The clustering approach can be roughly classified into partitioning method and hierarchical method. The partitioning clustering methods are CLARANS, PAM or K-means.

PAM algorithm

PAM is known as Partitioning around medoids or K-mediods. The main approach of this algorithm is to discover the k clusters from n items. First, randomly find the item which is called as medoids. Every object is clustered with Mediod which is similar. PAM algorithm follows the same steps that K-means algorithm will follow. PAM algorithms are based on the design of distance between data points which are frequently which compute by Euclidean metric.

K-means algorithm cannot handle categorical variable but Pam algorithm can handle it. PAM algorithm works similar to K-means rather than finding and updating the center of the cluster. PAM swaps all the data points in the cluster instead of judge the cluster.

Algorithm: Partition Around Mediods (PAM)

Input: i) K: enter a number of clusters should obtain.
ii) D: D is the data set which contains n objects i.e., the CSV file containing the extracted data

Output: It will display the k clusters based on an algorithm.
Steps:

(a) Clustering is applied to the CSV file which is extracted, in which have the attribute is called review.

(b) In the Review attribute, all words are separated by spaces these words are used to convert from string to vector form.

(c) Partitioned around medians (PAM) algorithm is used to find the similarity between the objects to perform clustering. In which it will randomly choose the k value as objects in D which is the original objects representative.

(d) Repeat and then assign all the remaining objects to cluster, by the closest object representative.

(e) Choose Orandom and calculate the whole cost and swap Oj with Orandom if S<0 to form the latest set of k object representative.

(f) Repeat the process until there will be no changes.

3.4. Classification

Classification is one of the data mining technique which is based on machine learning which can be used to sort the data entry in the site of the data which is interested in the predefined program.J48 is been used in this. The j48 algorithm is based on the decision tree. Classification is one of the processes of structuring a model for the set of reports which have a class label.

J48 algorithm

The splitting criteria of J48 are done for both numeric and character data. The numerical value can split based on the binary split. The character data can split to binary or any number of splits. The additional feature for J48 is for the last value, decision trees prune, the origin of rules, etc.

Algorithm: J48 Algorithm

Input: The Clustered data in CSV format.

Output: Analyses and classifies the reviews into positive words and negative words.
Steps:

(a) Dataset is given as input.

(b) For the Input, dataset calculates the Entropy and Information gain.

(c) By using splitting criteria which attribute will get higher information gain that attributes is made as the root node.

(d) Repeat the process and calculate for all the attributes and make it as a child node.

(e) Pruning means selecting a subtree that leads to the lowest test error rate. We can use cross-validation to determine the test error rate of a subtree.

(f) The tree can stop growing once the words are correctly classified.

J48 is called as Pruneable classifier tree since it is having the concept of Pruning. The pruning is having Post pruning and Pre pruning. In this, if the accuracy of the root node is more than the leaf node then it will remove leaf node. If the leaf nodes, are more accurate than leaf node then it will regenerate.

E. Dictionary Words

In Dictionary words it contains the List of Positive opinion words or the sentimental words which are having 2005 sentimental Keywords and the List of Negative opinion words or the Sentimental words which are having 4783 sentimental Keywords which are stored in two text files.
Example 1: Positive words= {abound, abounds, abundance, abundant, accessable , ……..}
Example 2: Negative words= { 2-faced, 2-faces,abnormal, abolish, abominable, ……..}

4. EXPERIMENT AND RESULT

The HTML document is fetched, parsed and finds the data which is contained in it. The class name of the element is specified to get the required HTML document. The extracted data of hotel web page is stored in the CSV file format. It contains six attributes and 109 instances. The six attributes are Hotel name, Reviews, Stars, Hotel Address, Hotel Link and Review Text and the 109 instances are the names of all the Hotels.

The Clustering uses PAM Algorithm on the Extracted data which is stored in the CSV file format. Based on PAM Algorithm it will form five clusters they are Cluster 0, Cluster 1, Cluster 2, Cluster 3 and Cluster 4. The graphical representation for Hotel Reviews is shown in the form of Pie chart in Fig -2 and in Table -1.

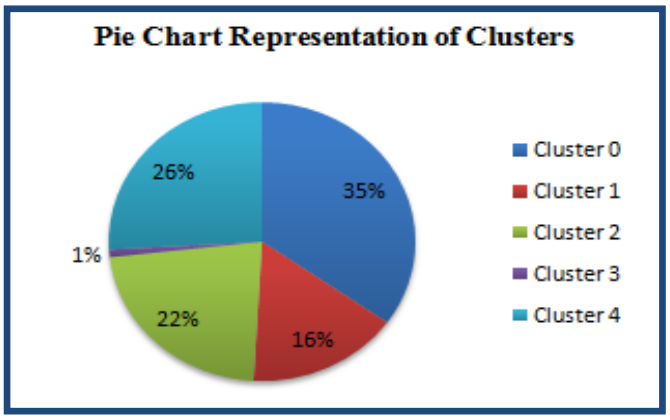

Fig -2: Pie Chart Representation of Clusters

In the extracted dataset there are 90 instances, the instances are clustered into 5 different clusters. Table -1 shows the number of instances and results in percentage in each cluster.

With the help of J48 Algorithm, the clustering data which is given as input will be classified and that data will be compared to the dictionary words and then give the Positive and Negative words which are obtained from all the hotel reviews on TripAdvisor.

All the Positive words are obtained based on the algorithm and count of the words are done for each hotel on the TripAdvisor website in one text file and the Negative words are obtained based on the algorithm and count of the words are done for each hotel in the TripAdvisor website in one text file.

Table -1: Representation of Clusters

Clusters	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Total Instances	39	18	25	1	29
Percentage of Instances	35%	16%	22%	1%	26%

The features are extracted from the reviews for all the hotels on the TripAdvisor website. The graphical representation for the count of the Positive and Negative words is shown in the form of Pie chart and table below.

Considering the example of one hotel, in which it is having the 12 Positive reviews and 2 Negative reviews. The graphical representation for the Hotel Reviews is shown in the form of Pie chart in Fig -3 and in Table -2.

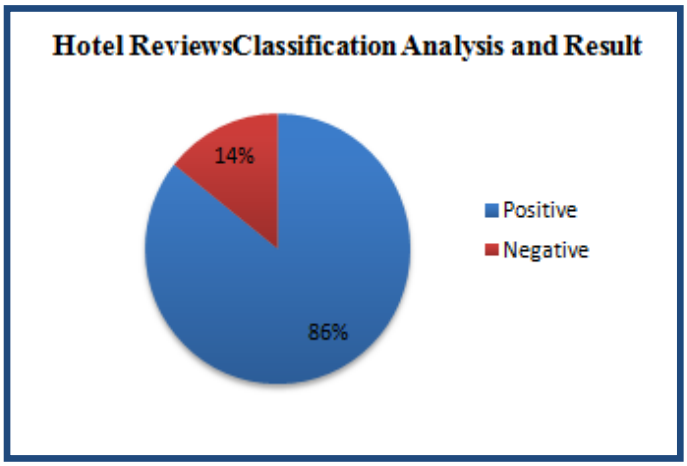

Fig -3: Pie Chart representation for Classification of Words

Table -2 represents the total number of words obtained and the percentage of it for the classification of words for the particular hotel.

Table -2: Result of Classification of Words

Classification of words	Positive words	Negative words
Total words	12	2
Percentage of words obtained	86%	14%

Word cloud represents the prominence of individual word classifying by highlighting words in dissimilar style. The word cloud representation of Positive words is shown in Fig. 4 and negative words in Fig -5.

The TripAdvisor is having many hotels in which top ten hotels name will display. It will display based on the reviews and the star ratings.

Fig -4: Word cloud of Positive words

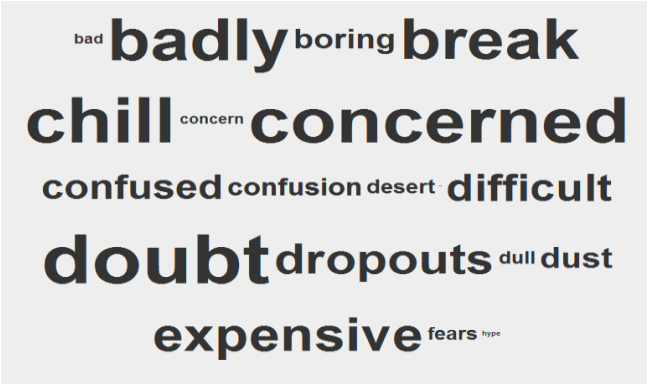

Fig -5: Word cloud of Negative words

5. CONCLUSIONS

Opinion Mining and Sentimental analysis on TripAdvisor for hotel reviews are achieved using keyword comparison and J48 Algorithm. The TripAdvisor is the data set it is only the base for extracting all details like Hotel name, number of reviews for each hotel, Star ratings, Hotel address and hotel link. The Hotel reviews are classified to find out the Positive words and Negative words and the count of the Positive and Negative words are calculated. The J48 algorithm is to classify Positive words and Negative words in reach review of all the hotels on the TripAdvisor website. And will also display Top ten hotels from the TripAdvisor website.

This application is restricted to only Bangalore location. The future work can carry out by considering various other locations like Chennai, Mumbai etc. Also, the hotels can be classified on basis of services, cleanliness, staff feedback and maintenance.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Prof N.R Shetty, Advisor, Dr. H.C Nagaraj, Principal, Nitte Meenakshi Institute of Technology and their Parents for giving constant encouragement and support to carry out research at NMIT.

REFERENCES

[1] Vijay B. Raut, D.D. Londhe, "Opinion Mining and Summarization of Hotel Reviews", Sixth International Conference on Computational Intelligence and Communication Networks, Pune, India,2014.

[2] Betul Dundar, Suat Ozdemir, Diyar Akay "Opinion Mining and Fuzzy Quantification in Hotel Reviews", IEEE TURKEY, 2016

[3] Apurva Juyal*, Dr. O. P. Gupta " A Review on Clustering Techniques in Data Mining" International Journal of

Advanced Research in Computer Science and Software Engineering Volume 4, Issue 7, July 2014

[4] Isaac B. Muck, Vasil Hnatyshin, Umashanger Thayasivam "Accuracy of Class Prediction using Similarity Functions in PAM" Glassboro, 2016 IEEE.

[5] Sunita Joshi, Bhuwaneshwari Pandey, Nitin Joshi "Comparative analysis of Naive Bayes and J48 Classification Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 12, December 2015

[6] Gaganjot Kaur, Amit Chhabra "Improved J48 Classification Algorithm for the Prediction of Diabetes" International Journal of Computer Applications Volume 98 – No.22, July 2014.

[7] Tri Doan and Jugal Kalita" Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning" 15th IEEE International Conference on Machine Learning and Applications 2016

[8] Hongting Li, Qinke Peng, Xinyu Guan "Sentence Level Opinion Mining of Hotel Comments", Proceedings of the IEEE International Conference on Information and Automation Ningbo, China, August 2016.

[9] Gopi Gandhi, Rohit Srivastava " Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms" International Journal of Computer Applications Volume 87 – No.9, February 2014.

[10] Walaa Medhat, Ahmed Hassan, Hoda Korashy "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal (2014) 5.

[11] Prerna Kapoor1, Reena Rani "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning" International Journal of Engineering Research and General Science Volume 3, Issue 3, May-June, 2015

[12] http://ptrckprry.com/course/ssd/data/positive-words.txt

[13] http://ptrckprry.com/course/ssd/data/negative-words.txt