DESIGN AND IMPLEMENTATION OF WEB APPLICATION FOR REVIEW CLASSIFICATION

Siddhi Pardeshi[1], Suyasha Ovhal[2], Pranali Shinde[3], Manasi Bansode[4], Anandkumar Birajdar[5]

[1-4]Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India
[5]Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India

---***---

Abstract - *As the technology is growing rapidly everyone is expressing their views in many websites. The rating of any online commercial site is highly dependent on the opinion of its users. Not only commercial sites nowadays every type of business-like hotels, banks, shops, malls etc are available on the internet. They all seek online ratings from their consumers. Consumers put their opinion according to how they experienced the service. The proposed work is for the analysis of hotel reviews. Consumers who visited certain hotels put their experience on the hotel website. On the basis of the reviews we can analyze user experience for the hotels and compare with other hotels. The proposed research work is implemented through a Machine learning algorithm named Random Forest Algorithm.*

Key Words*: *Machine Learning, Random Forest, Review Classification

1. INTRODUCTION

In present days a company, a business organization or service-based companies which requires feedback from its customers. Increase in expansion of company will be providing more number of services and products. So, as to increase this the organization must bother about the reviews and ratings given by its users. The service-consumers can mention their feelings and reviews on online-portals. By performing review classification we can predict the goodness of that particular organization. The Customer's sentiments regarding to a hotel depends upon the facilities he/she got from that hotel, just like cleanliness, location of the hotel, services provided by the hotel like free wi-fi, multilingual staff, bar/lounge, babysitting rooms, wheelchair etc.

The sentiments of customers can be expressed in the form of excellent, good, average, poor, terribleetc. Generally, customers want to express their feelings also with these rating and review values. Throughout our paper we will explain how we implemented a web application for booking tables and rooms in hotels and also classifying the reviews entered by user into positive and negative reviews.

2. LITERATURE REVIEW

In 2017, Zeenia Singla, Sukhchandan Randhawa and Sushma Jain performed sentiment analysis of customer product reviews [7]. In this they collected the E-commerce Amazon dataset, performed pre processing steps like stopword removal, punctuation removal, stemming etc. and selected three features out of six using feature selection methods. Then they performed sentiment analysis and calculated the polarity of each review. Based on the calculated polarity, they divided the reviews into two parts as positive and negative reviews by appending pos/neg words to each review. They used the undersampling technique to balance the unbalancing of positive and negative reviews. After that they used three classification models as Naive Bayes, Decision Tree and SVM, out of which SVM had given the best accuracy when tested for 10 fold cross validation.

In 2018, Andreea Salinca used the Yelp Challenge Dataset with 1.6 million reviews in which the main focus was on star rating and business raw text reviews [8]. Reviews having star rating above 4 were considered as 'positive' and below 3 were considered as 'negative' , while rating equal to 3 were eliminated. The dataset was splitted (80% for training and 20% for testing) and preprocessing techniques were used to extract a set of features. Preprocessing techniques like stemming, stop words removal were performed. Custom dictionary was built in the first approach and lexical analysis of text reviews are performed in the second approach. They performed the classification process using 3-fold cross validation for evaluating the accuracy. Then different machine learning classifiers like Logistic Regression, Naive Bayes, Stochastic Gradient Descent (SGD), and Linear Support Vector Classification (SVC), classifiers were used to classify review and to extract features. Author obtained 94.4% accuracy using SVC and SGD while naive Bayes and logistic regression had the worst results.

In 2016, Kimitaka Tsutsumia, Kazutaka Shimada a and Tsutomu Endoa used movie review dataset for the classification based on multiple classifiers [9]. They preprocessed the dataset by stopword removal and stemming and had done the feature selection. Then they proposed a method which consisted of three classifiers as-

SVM, Maximum Entropy and Score Calculation and the output of these classifiers was integrated with the help of naive voting and weighted voting methods. They also used SVMs again in the integration process which produced the best performance. The main focus of this paper was that the multiple classifiers outperformed the single classifiers.

In 2018 [10]. for summarizing the opinions of the reviews which were mentioned by the users Abdi etal offered a machine learning technique. The method consisted of merging multiple kinds of features into a unique feature set for accurate modelling of the classification model. Thus, an investigation on performance was done for four best feature selection models for accomplishing the best performance and seven classifiers were selected for choosing the applicable feature set. Hence recognised an effective machine learning model. The methods which were suggested was implemented in different datasets. The output thus specifies that for feature selection approach IG come out to the best and for classification of the data SVM-based approach enhances the performance.

3. OVERVIEW OF SYSTEM IMPLEMENTATION:

A. SYSTEM DESIGN

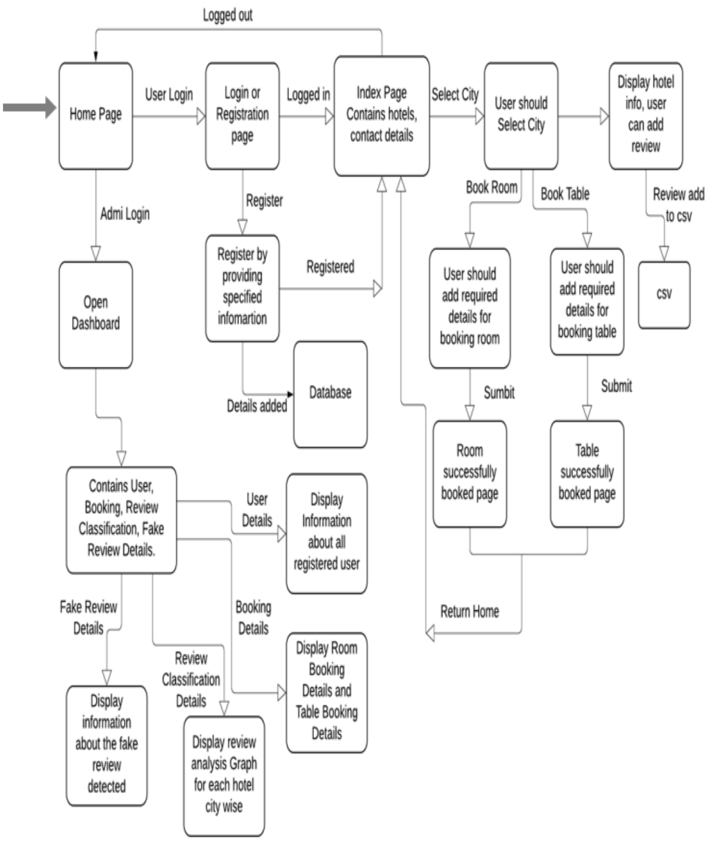

Logged out
Home Page
User Login
Login or Registration page
Logged in
Index Page Contains hotels, contact details
Select City
User should Select City
Display hotel info, user can add review
Admi Login
Register
Book Room
Book Table
Review add to csv
Open Dashboard
Register by providing specified infomartion
Registered
User should add required details for booking room
User should add required details for booking table
csv
Database
Details added
Sumbit
Submit
Room successfully booked page
Table successfully booked page
Contains User, Booking, Review Classification, Fake Review Details.
User Details
Display Information about all registered user
Fake Review Details
Booking Details
Return Home
Review Classification Details
Display Room Booking Details and Table Booking Details
Display information about the fake review detected
Display review analysis Graph for each hotel city wise

Fig 3.1. System Architecture Design

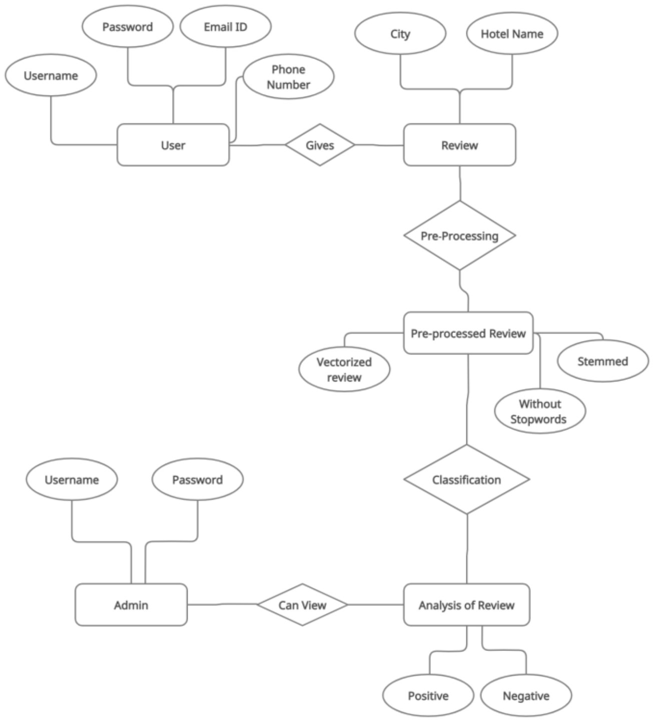


Fig 3.2. E-R Diagram

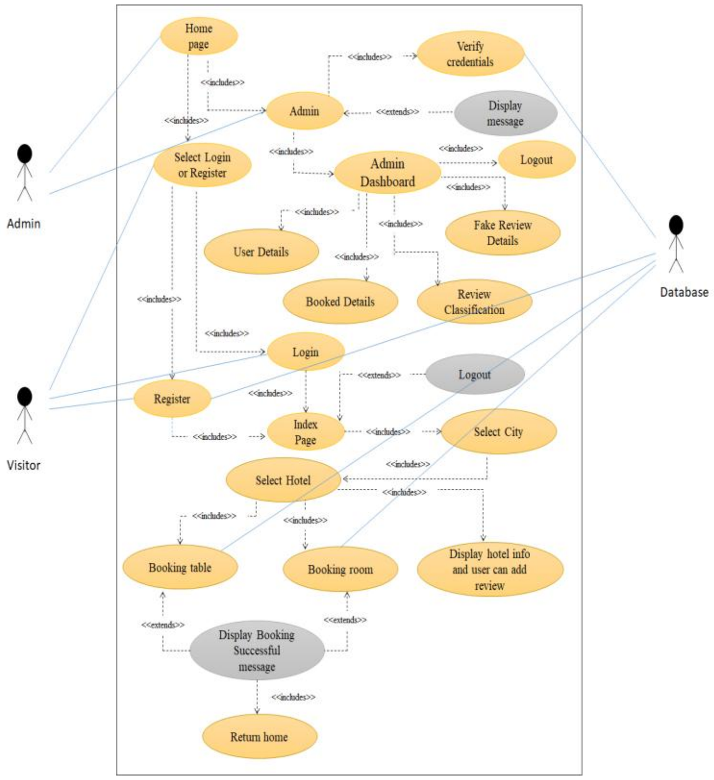


Fig 3.3. Use Case Diagram

B. SYSTEM IMPLEMENTATION

To accomplish our goal, we trained our model on a hotel review dataset. Above system diagram (Figure no.3.1) represents the basic flow of our system. Users need to first register, to avail the features provided by our web application. User details are inserted into MySQL database. User is then directed to the main page where he can hand-pick between various cities. For each hotel compendious details are given on web application. Further user can choose among provided hotels to either book a room or book a table. User details are inserted into MySQL .

Users can refer reviews provided for each hotel. Users can add their reviews for desired hotels according to their experience which are then added to the csv file for analysis. Modules then parses the reviews and classifies them as positive and negative.

Our web application is managed by Admin. Registered user details are displayed on the admin dashboard. Admin can also go through rooms or table booking details. Result of review classification for each hotel and their details are displayed on dashboard through API.

Libraries Used:

Python offers a reserve of inbuilt libraries. Libraries contain modules which provide various functionality. Libraries used in our web application are as follows:

·**Numpy** : The NumPy library is used for array manipulation and scientific computing in python.

·**Pandas** : It provides structures and operations for manipulating numerical tables and time series.

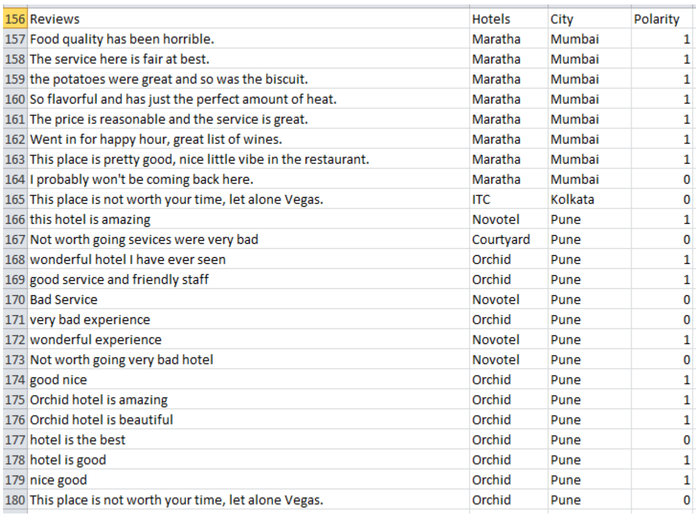
·**Matplotlib** : Matplotlib is a comprehensive library for creating animated, static and interactive visualizations in Python.

·**Flask** : Flask is a micro web framework written in Python and provides tools to develop web applications. We have used flask framework to develop a web application for hotel booking.

·**MySQLdb** : MySQLdb provides users the functionality to connect and perform various operations on MySQL database.

Storage Information

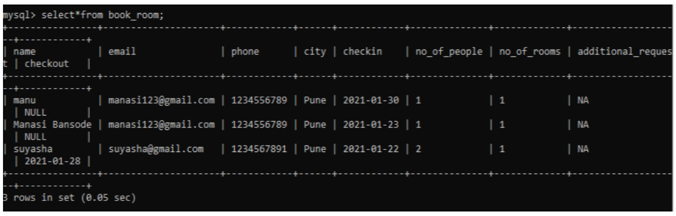
1)CSV : CSV (Comma Separated Values) format is the most popular import and export format for databases and spreadsheets.CSV files is used in our web app for storing and processing reviews for their classification.CSV file is fed to machine learning module which classifies the reviews entered by user on web application and stores the result in result CSV file. The sample data from csv file which has results of review classification is shown below.



156	Reviews	Hotels	City	Polarity
157	Food quality has been horrible.	Maratha	Mumbai	1
158	The service here is fair at best.	Maratha	Mumbai	1
159	the potatoes were great and so was the biscuit.	Maratha	Mumbai	1
160	So flavorful and has just the perfect amount of heat.	Maratha	Mumbai	1
161	The price is reasonable and the service is great.	Maratha	Mumbai	1
162	Went in for happy hour, great list of wines.	Maratha	Mumbai	1
163	This place is pretty good, nice little vibe in the restaurant.	Maratha	Mumbai	1
164	I probably won't be coming back here.	Maratha	Mumbai	0
165	This place is not worth your time, let alone Vegas.	ITC	Kolkata	0
166	this hotel is amazing	Novotel	Pune	1
167	Not worth going sevices were very bad	Courtyard	Pune	0
168	wonderful hotel I have ever seen	Orchid	Pune	1
169	good service and friendly staff	Orchid	Pune	1
170	Bad Service	Novotel	Pune	0
171	very bad experience	Orchid	Pune	0
172	wonderful experience	Novotel	Pune	1
173	Not worth going very bad hotel	Novotel	Pune	0
174	good nice	Orchid	Pune	1
175	Orchid hotel is amazing	Orchid	Pune	1
176	Orchid hotel is beautiful	Orchid	Pune	1
177	hotel is the best	Orchid	Pune	0
178	hotel is good	Orchid	Pune	1
179	nice good	Orchid	Pune	1
180	This place is not worth your time, let alone Vegas.	Orchid	Pune	0

Table 3.1 : Review analysis

2)MySQL : MySQL is an open-source relational database management system (RDBMS).MySQL is used to store, retrieve and maintain registered users, room and table booking details. These details are managed by admin. Below a sample table data stored in MySQL for room booking is shown.



```
mysql> select*from book_room;
+----------------+---------------------+------------+------+------------+--------------+-------------+-------------------
--+------------+
| name           | email               | phone      | city | checkin    | no_of_people | no_of_rooms | additional_reques
t | checkout   |
+----------------+---------------------+------------+------+------------+--------------+-------------+-------------------
--+------------+
| manu           | manasi123@gmail.com | 1234556789 | Pune | 2021-01-30 | 1            | 1           | NA
  | NULL       |
| Manasi Bansode | manasi123@gmail.com | 1234556789 | Pune | 2021-01-23 | 1            | 1           | NA
  | NULL       |
| suyasha        | suyasha@gmail.com   | 1234567891 | Pune | 2021-01-22 | 2            | 1           | NA
  | 2021-01-28 |
+----------------+---------------------+------------+------+------------+--------------+-------------+-------------------
--+------------+
3 rows in set (0.05 sec)
```

Table3.2 : Book room table

C. OVERVIEW ON ALGORITHMS AND TECHNIQUES USED

1. Dataset

The dataset which we have used contains hotel reviews collected from kaggle [13] which consists of 515739 rows and 16 columns. Then data is passed for preprocessing and cleaning of data.

2. Data Pre-processing

Before passing the data directly to the model it should be pre-processed. The pre-processing of the model consists of replacing missing review info by empty strings then again replacing that empty string with NAN values and then dropping them all.

3. Data Cleaning

For cleaning the data we have first removed all the stop words by using Natural Language Toolkit (nltk) library. Then further we have used Stemming for producing morphological variants of a root word using Porter

Stemmer. A Stemming algorithm reduces a word into its root word for example “chocolaty”, “Choco”, “chocolates” to root word, “chocolate” and “retrieves”, “retrieved” ,“retrieval” reduced to the stem “retrieve”. Then we have removed the unwanted words like ['room', 'staff', 'locat', 'hotel', 'breakfast', 'bed', 'shower', …..] which do not contribute to the review classification.

4. Model Selection

After all the pre-processing and cleaning techniques the data is then converted to vector form using Count vectorizer and passed to the Random Forest Classifier model for classification. Then the model is trained and it is stored as a pickle file with .pkl extension. Further this pickle file is used for performing review classification on testing data.

5. Random Forest Algorithm

Random Forest is a supervised machine learning algorithm which is used for both regression problem and classification problem. Random Forest is the most adaptable and uncomplicated to use. A Forest consists of trees and here the more number of trees a forest has, the more vigorous a forest is. Random forest generates many decision trees on randomly selected data. It then fetches predictions from every tree created and chooses the best optimum solution by voting. Random forest is also a good indicator of feature importance. Random Forest has many applications, like for creating text and image classification models , a commendation model, feature selection etc.

Random Forest Classifier is also mostly used in semantic sentiment analysis. In our system the classifier takes two individual classes of positive words and also negative words. Then the model will construct a conditional density table which will be based on frequencies of positive words and the negative words already obtained by pre-processing of data. And at the end model will return the classified result in the out.csv file.

Here general technique of bootstrap aggregating or bagging is used to tree learners for training Random Forests. Training set is given as X=x1, x2,…, xn. and the responses will be Y= y1, y2,…, yn. By bagging repeatedly for B times will select a random sample and that will be replaced by the training set and will fit trees to these samples.

For b=1,2…, B.

1. Sample after the replacement, n training examples from X,Y will call Xb,Yb.

2. Then training a classification tree for Xb,Yb.

Predictions for all unseen samples x’ made by taking majority vote for classification problem statement.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Equation for above explanation

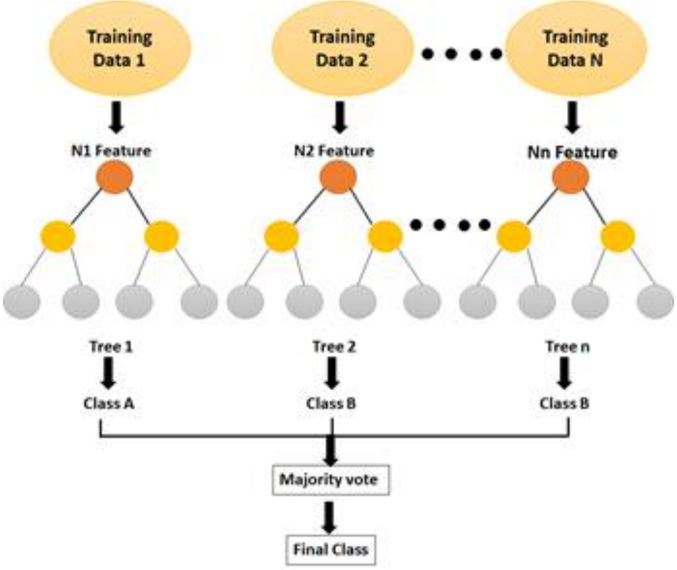


Image 3.4 : Random forest algorithm model

D. EVALUATION AND RESULT

Following are some snapshots of our web application implemented using Flask web framework.

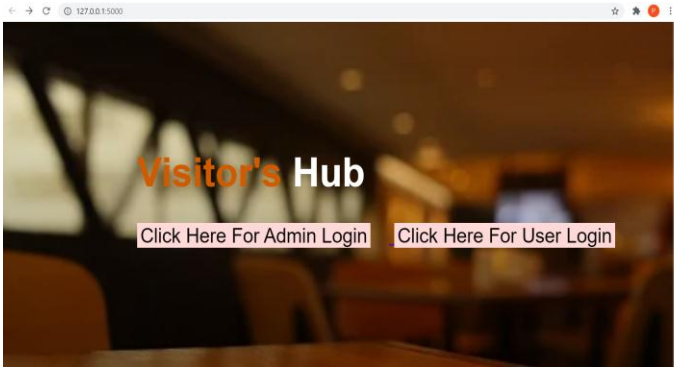
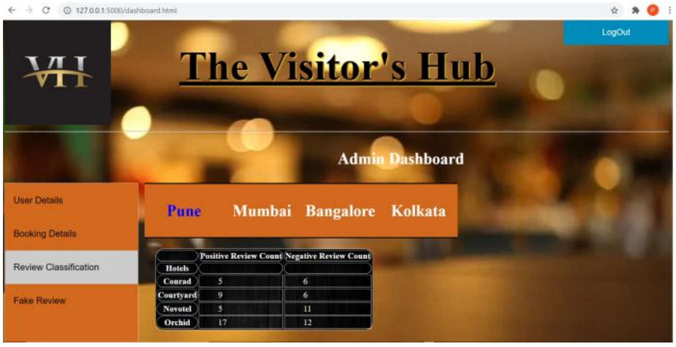
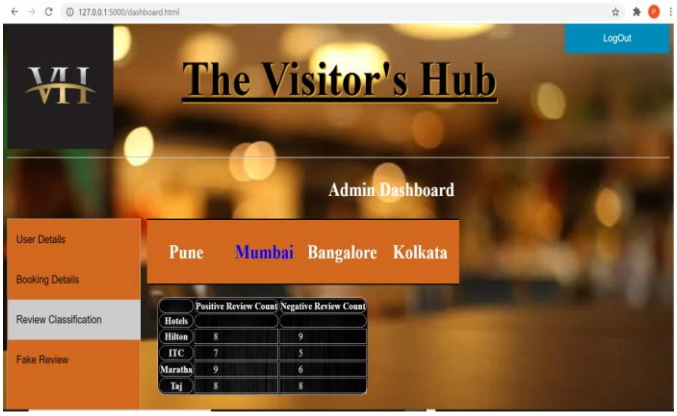


Image 3.5 : Home Page



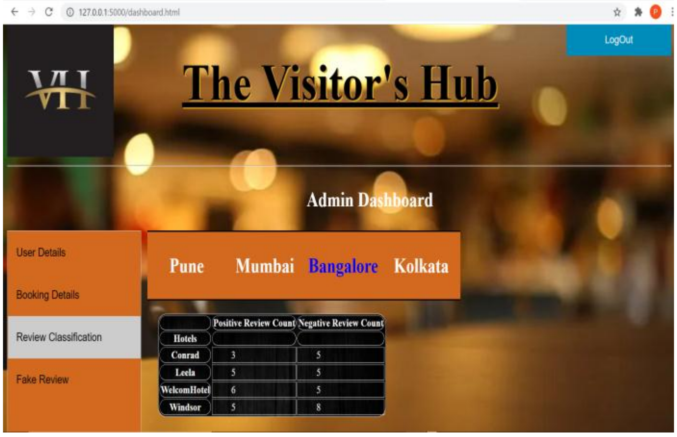
	Positive Review Count	Negative Review Count
Hotels		
Conrad	5	6
Courtyard	9	6
Novotel	5	11
Orchid	17	12

Image 3.6 : Admin Dashboard (Hotel Review analysis for Pune City)



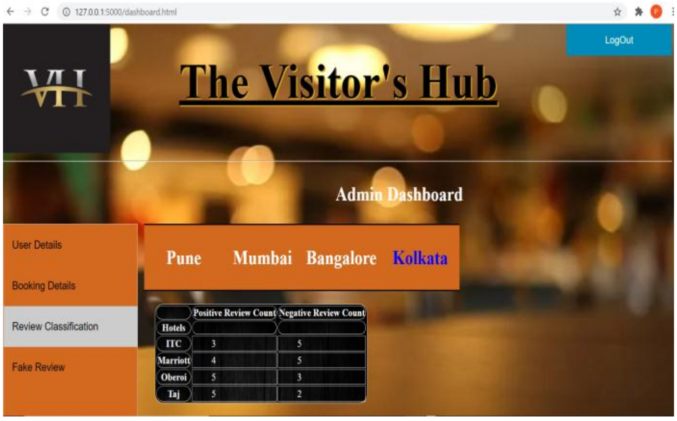
	Positive Review Count	Negative Review Count
Hotels		
Hilton	8	9
ITC	7	5
Maratha	9	6
Taj	8	8

Image 3.7 : Admin Dashboard (Hotel Review analysis for Mumbai City)



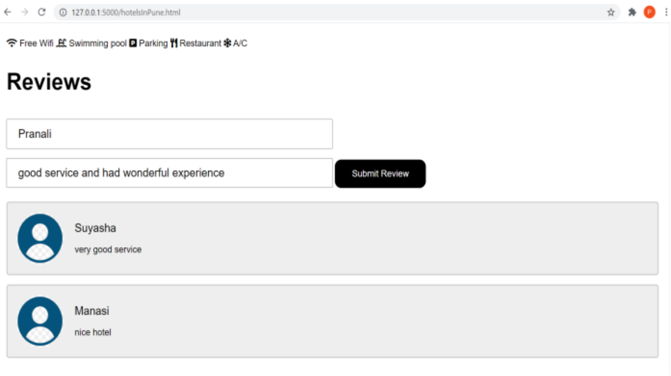
	Positive Review Count	Negative Review Count
Hotels		
Conrad	3	5
Leela	5	5
WelcomHotel	6	5
Windsor	5	8

Image 3.8 : Admin Dashboard (Hotel Review analysis for Bangalore City)



	Positive Review Count	Negative Review Count
Hotels		
ITC	3	5
Marriott	4	5
Oberoi	5	3
Taj	5	2

Image 3.9 : Admin Dashboard (Hotel Review analysis for Kolkata City)


127.0.0.1:5000/hotelsInPune.html
Free Wifi Swimming pool Parking Restaurant A/C
Submit Review

Reviews

Pranali

good service and had wonderful experience

Suyasha
very good service

Manasi
nice hotel

Image 3.10 : User given reviews

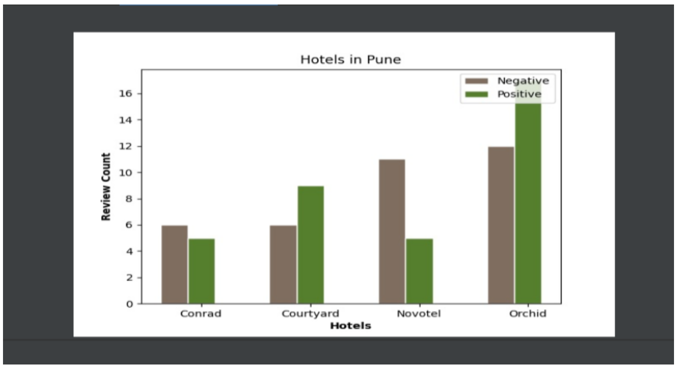


Image 3.11 : Hotel Review analysis for Pune City graph

4. CONCLUSION

A web application was successfully built to allow customers to book rooms or tables in hotels and give reviews about their experience. Also review classification for the user given hotel reviews was carried out. The reviews were classified as positive reviews which include words like happy, amazing, tasty, nice, good, pretty etc and as well as negative reviews which include words like bad, disgusting, sad, and disappointed, etc. The whole point of the analysis is to provide suitable recommendations to the customers to select the best available hotel option. It also helps the hotel owners to know what customers think about their hotel facilities.

5. FUTURE SCOPE

Review classification divides reviews into two categories as positive reviews and negative reviews. But reviews given by the user might be fake. Thus, in future, in order to know whether the review is fake or genuine we can perform fake review predictions.

6. ACKNOWLEDGEMENT

We express immense gratitude to our Seminar Guide Prof. Anandkumar Birajdar for his support and encouragement. We also especially thank him for his useful suggestions and having laid down the foundation for the success of this work.

We would also like to thank our Research & Innovation coordinator Prof. Dr. Swati Shinde and Project Coordinator Prof. Smita Khairnar for their genuine support, guidance, and assistance from early stages of the seminar. We would also like to express our gratitude to Prof Dr. K. Rajeswari, Head of Computer Engineering Department for her unwavering succour during the whole seminar activity. We would also like to thank all the web communities for enriching us with their enormous information.

7. REFERENCES

[1]Fake Review Detection using Data Mining, Md Forhad Hossain, Missouri State University, Mdforhad08@live.missouristate.edu, Summer 2019.

[2] An Empirical Study on Detecting Fake Reviews Using Machine Learning Techniques, Elshrif Elmurngi, Abdelouahed Gherbi, Department of Software and IT Engineering École de Technologie Supérieure Montreal, Canada, The Seventh International Conference on Innovative Computing Technology (INTECH 2017).

[3]https://www.kaggle.com/athoul01/predicting-yelp-ratings-from-review-text

[4]Fraud Detection in Online Reviews using Machine Learning Techniques,Kolli Shivagangadhar, Sagar H, Sohan Sathyan, Vanipriya C.H(2017)

[5]Classifiers Ensemble for Fake Review Detection,Harish Baraithiya, R. K. Pateriya(2019)

[6] Fake Review Detection using Opinion Mining, Dhairya Patel, Aishwarya Kapoor, Sameet Sonawane, 2018.

[7] Sentiment Analysis of Customer Product Reviews Using Machine Learning, Zeenia Singla, Sukhchandan Randhawa, Sushma Jain,2017.

[8] Business reviews classification using sentiment analysis, Andreea Salinca, 2018.

[9]Movie Review Classification Based on Multiple Classifiers, Kimitaka Tsutsumia, Kazutaka Shimada a and Tsutomu Endoa, 2016.

[10]AsadAbdi, Siti MariyamShamsuddin, ShafaatunnurHasan, and JalilPiranMD, "Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment", Expert Systems with Applications, vol. 109, pp. 66-85, 1 November 2018.

[11]Research Paper Classification Systems Based on TF-IDF and LDA schemes, Sang‑Woon Kim and Joon‑Min Gil,2019.

[12]Text Classification Algorithms: A Survey,Kamran Kowsari ,Kiana Jafari Meimandi,mojtaba Heidarysafa,Sanjana Mendu, Laura Barnes and Donald Brown,2019.

[13]https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe