International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 22 (2018) pp. 15830-15835

Machine Learning Based Comprehensive Analysis of Hospitality Industry in the State of Karnataka

Shruthi C G
Department of Computer Science and Engineering
Dr.Ambedkar Institute of Technology
Banglore-56, Karnataka, India.

Gowrishankar S.
Department of Computer Science and Engineering
Dr.Ambedkar Institute of Technology
Banglore-56, Karnataka, India.

Abstract

With the increase in the evolution of the internet, there is a millions of data provided by the users in the form of reviews available in online websites of various fields. One among them is a tourism websites called Tripadvisor.in. The central aim of this paper is to scrap the reviews and other review related information from this website, store this dataset in the MySql database and then load this dataset into pandas for future analysis of data, analyse the hotels reviews in the districts of Karnataka using various machine learning technologies like Data Visualization of the retrieved data in a meaningful graph through the data available in the database using python libraries called Matplotlib and Seaborn. Representing any data in the visual form conveys the user to easily understand what is actually happening with the data rather than text or tables. We can also visualize the data in many ways one among them is Word Cloud, where we can find the words frequencies in which word repeated maximum is in the larger font appearance and minimum repeated word in the small font in visual form. This paper also proposes fast and accurate sentiment classification using an Enhanced Naive Bayes mode where we have also explored whether a given sentence is positive or negative or neutral and how confidently a given data is positive, negative and neutral and finally we have predicted the best hotels based on their star ratings. This has been done through the use of machine learning algorithm called Logistic Regression.

Keywords: Web Scraping, Database Server, Data Visualization, Word Cloud, Sentimental Analysis, Prediction.

I. INTRODUCTION

whenever a person has to make a decision about particular place or product, he or she must have some vague idea about it. Here we opted for the tourism field and to know the opinion of the other peoples about that particular place or their experience However there is an enormous data generated online in the form of reviews. Here comes where we to land up, Now we considering internet as the central resource for opinion evaluation. Following one of the popular tourism website called tripadvisor.in. There we mainly check for user comments, star ratings and their reviews about the hotels[1]. This data is scraped from this website and retrieved data is visualized in different ways like Graphs, Word cloud. Sentimental Analysis and Prediction using machine learning algorithms like Enhanced Naive Bayes[7] and Logistic Regression algorithms. in order to show the best hotels among the available hotels in various districts of Karnataka as discussed[8].

Our all this process is mainly a part of Data science that uses methods, scientific ways, algorithms to extract data in the different forms for data processing. and Machine Learning uses mathematical analytics and techniques to provide computers the flexibility to learn with information, while not being expressively programmed Machine learning is closely associated with (and usually overlaps with) machine statistics, that conjointly focuses on prediction creating through the employment of computers.

In this paper, we analyse the seven steps. Starts from Scraping the data from Website called tripadvisor.in[1]. Once data is retrieved need to store that data for further reference, then we have gone with the storing of retrieved dataset in the MySql Database server[2]. Loading the dataset to Pandas, we have used pandas to represent the complex data in the expressive and flexible format, through this we can parse, load and transfer data[3] that Visualizing the data in the form of meaningful graphs using Matplotlib and Seaborn[4]. Visualizing the words frequency using Word cloud to check the maximum and minimum repeated words in visual form[5]. Performing the sentimental analysis using Naive Bayes algorithm[7]. Prediction of the best in some districts of Karnataka using one of the prediction algorithm available in machine learning called Logistic Regression[8].

In the second section we are going through the literature survey which gives the details and knowledge which made our work into practical. third section called methodologies which gives information about how actually implementation flow works in a step by step starting from web scraping followed by data server, data visualization, word cloud sentimental analysis and finally the prediction. and at last that is fourth section is the results and discussions the outputs of all the step that we have implemented in our project.

II. LITERATURE SURVEY

Web scraping or web data extraction is nothing but extraction of the required information from the online websites of any field we are interested in, extracted dataset from the web is stored in this database for further retrieval and analysis. Data visualization is done in the form of some meaningful graphs that contains some useful information in it[1].

Authors in paper[5] have visualized the data in terms of word frequency check that is in visual from though the word cloud "Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Word cloud Visualization", we need data representation in visual form that is easier to understand than text and tables. We can also visualize the data in many ways one among them is Word cloud. It is a visual representation of the frequency of words that occurs in text books or websites. Font size specifies the occurrence frequency of a word: the bigger the font of the word size, the greater the word frequency is and in opposite, the smaller the font of the word size, the lesser the word frequency is.

"Sentiment Analysis of Movie Review Data Using Senti-Lexicon Algorithm", is executed by the authors in papers[6]. The central aim of this research paper is to perform sentiment analysis on movie review data. They have proposed the Senti-lexical algorithm to find the polarity of a review as positive, negative or neutral. They have also proposed a method to handle words which have negation effect on the reviews and the role of emotions is also discussed.

"Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model", We observed that a combination of methods like effective negation handling, word n-grams and feature selection by mutual information results in a significant improvement in accuracy is implemented by excellent scientists in paper[7].

"Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.)". In other words, the logistic regression model predicts P(Y=1) as a function of X. Susan Li, Senior Data Scientist, Kognitiv Corp in September 7, 2017. He went through an example explaining the working of logistic regression for one of the banking sector[8] related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y). The dataset can be downloaded from certain database. Here he said the dataset comes from the UCI Machine Learning repository. The dataset provides the bank customers' information which includes 41,188 records and 21 fields.[8].

III. METHODOLOGIES

Analysis of Karnataka tourism industry in terms of hotel review user data using machine learning technologies in order to produce best hotels for the tourists.

We are trying to solve this by scraping one of the popular tourism website called Tripadvisor.in, retrieving the relevant data from this website using python libraries called BeautifulSoup and Urllib for fetching urls. Retrieving the data, storing this data in database, visualization of the hotels related data in the form of meaningful graphs using some of the python libraries like Seaborn and Matplotlib, visualization of the word frequencies through Word Cloud, sentimental analysis of the user reviews nothing but the user emotion towards the hotel in terms of positive, negative or neutral using sentimental analysis algorithms called fast and accurate sentiment classification using Enhanced Naive Bayes model[7], and predicting the best hotels in certain districts of Karnataka through prediction algorithms called Logistic-Regression algorithm[8]. The architecture above gives us the clear picture of flow how the work is been done from initial step to final step.

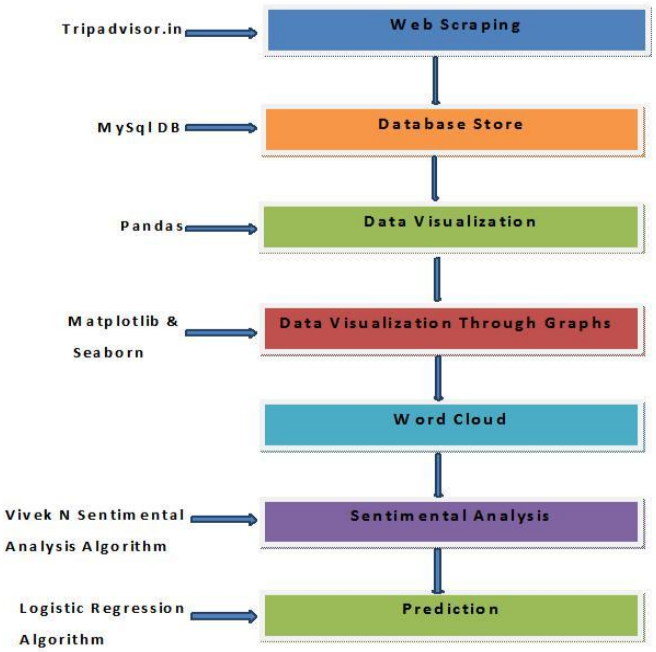

Fig.1 Flow diagram of proposed work

Before showing our emotion towards any product or place we need to have some idea about it , here we are talking about the tourism industry in Karnataka especially hotels. We can't estimate which hotel is good to go or bad to stay, so there are many online websites which gives us the reviews of the hotels given by the people who has visited it or experienced it before. We have selected one of the popular tourism site available online called Tripadvisor.in.

First and foremost is web scraping that defines retrieving the relevant data present in the website for our reference. We scrap the data using the python tool called Beautiful Soup of version4 and with the help of universal library called Urllib for fetching some of the hotels related url's by writing a piece of code in python.

The data scraping starts from logging on to the Tripadvisor.in website, search for Karnataka, then to popular destinations in Karnataka, here u find all the districts of Karnataka, After that we are entering into each districts hotels and their review pages. Our goal is to extract list of hotels in various districts of Karnataka, names of the hotels, address, total reviews including individual user reviews of each hotel. User comments and their experience in the form of description and their star ratings. And some details of the persons who has given the review, You can see how we have done the scrapping in a flow format below[1].

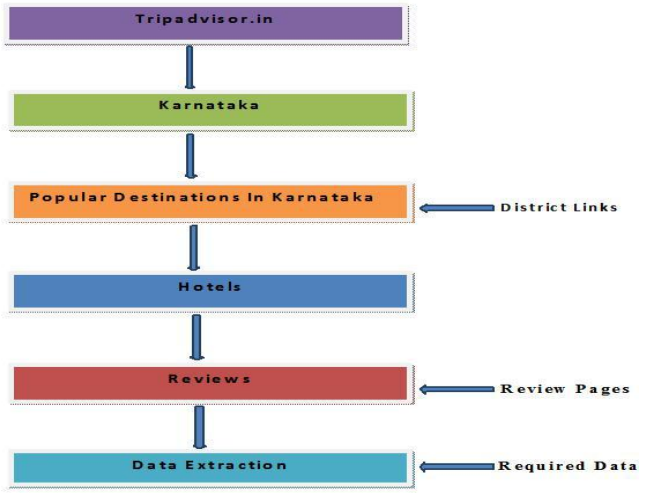

Fig.2 Flow diagram of web scrapping

The information is extracted for further analysis or processing. Second step is, once the data is extracted we need to store it in some place and here comes the concept of MySql database server. We have created 4 tables as listed in Fig 3, we developed a relationship between these tables, and storing the extracted data in the respective tables namely. Category table as specified in the database scheme. The Person Details table as represented in the scheme, with parameters and dataset stored in the Person Details table namely, Person Id, Location from where he has given the review and Date on which he posted the review.

Individual Reviews table with parameters and dataset namely, Review Id, Title of the Review, Review Description, Star Ratings, Person Id who has given that review and along with the Hotel Id and the Overall Review Details table as represented in the scheme, the parameters and dataset stored in the Overall Review Details namely, Hotel Id, Hotel Name, Hotel Address, Total Reviews Count, distribution of 100 percent ratings into Excellent, Very good, Average and Poor.

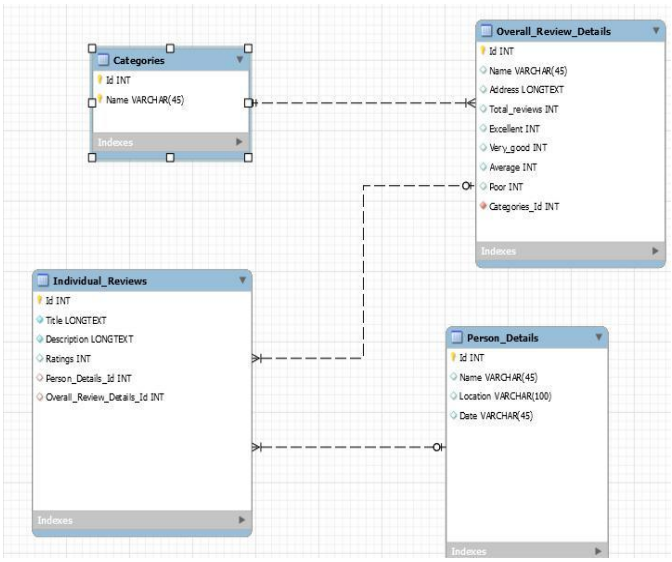

Fig.3 Database relational schema of the hotels data

Once the data is stored in MySql Database Server, we loaded our dataset to pandas for providing flexible and expressive data structures and now comes the concept of visualization, once the data is loaded into pandas, we do the data visualization using python libraries called Matplotlib and Seaborn. Each plot presents data in a different way, it is good to keep in mind that visualization is a blend of art and science.

Now it's time to visualize data in terms of Word Cloud that is we find the maximum repeated words in the reviews given by the users. We need data to be represented in visual form that is easily understandable than text and tables. That provides a visual representation of words frequency Font size specifies the occurrence frequency of a word: the bigger the font of the word size, the greater the word frequency is and in opposite, the smaller the font of the word size, the lesser the word frequency is.

Next we are doing the sentimental analysis to know the emotion of the person towards the place or product, to check whether the given review is a positive sentiment or negative sentiment or it is neutral we have used one of the Sentimental Analysis algorithm from Machine Learning called Vivek Narayanan sentimental analyzer Algorithm Using an Enhanced Naive Bayes Model. "Naïve Bayes is a very simple probabilistic model that tends to work well on text classifications and usually takes orders of magnitude less time to train when compared to models like support vector machine"[7]. We also show how confidently it is positive, negative and neutral.

Our last and final work is to show the users the best hotels in a particular district based on the review ratings given by the already experienced tourists or users. We have done this with the help of prediction algorithms available in the machine learning field, we preferred one of the algorithm called Logistic Regression algorithm for prediction of best hotels in a some district of Karnataka. In our paper we have predicted the best hotels of 2 districts like Banglore and Kodagu (Madikeri).

"Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.)". In other words, the logistic regression model predicts P(Y=1) as a function of X[8].

IV. RESULTS AND DISCUSSIONS

A piece of python code performs all this scraping dynamically and parses the review page links of hotels in each districts of Karnataka and extracted data gets loaded into the MySql database server. Now the following data that is Category table, The Person Details with parameters like Person Id, Location of the person, Date on when he posted the review.

Individual Reviews table with parameters like Review Id, Title review in short, Review Description, Star Ratings, Person Id and Hotel Id and the Overall Review Details with attributes Hotel Id, Hotel Name, Hotel Address, Total

Reviews Count, distribution of 100 percent ratings into Excellent, Very good, Average and Poor. The successive data is stored in the attributes specified in the tables in Fig 3.

Each plot is presenting us a meaningful data about the hotels. it is good to keep in mind that visualization is a blend of art and science. For this we have used some of the python libraries called Matplotlib, Seaborn, NumPy which mainly provides array objects, matrices and so..on, We use this library in our paper since we are visualizing the hotels in different districts of Karnataka.

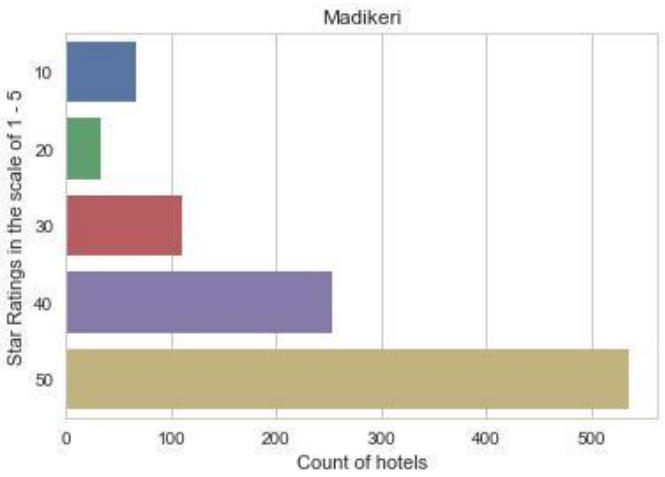

Fig.4 Star Ratings against number of hotels in Madikeri(kodagu)

Fig.4 gives the star ratings in the scale of 1 to 5 against the number of hotels in Madikeri. We have done the visualization of hotel star ratings among the overall available hotels in Madikeri. Here on graph, x-axis shows the number of hotels star ratings in the scale of 1 to 5 in Madikeri and y-axis represent one to five star ratings.

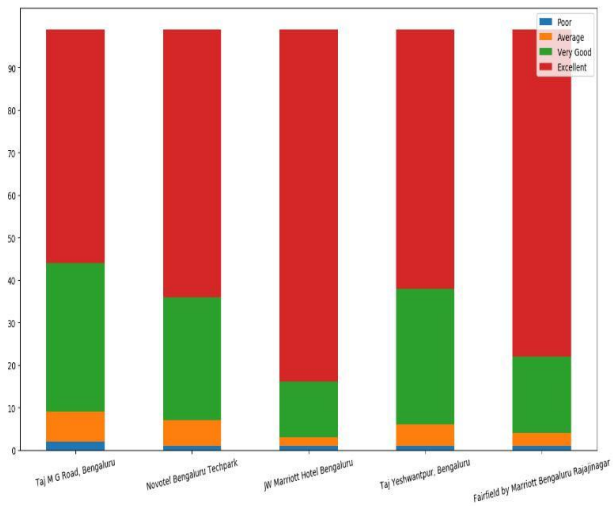

Fig.5 Distribution Of excellent, very good, average, and poor ratings for Banglore Hotels

The above Fig.5 gives the Banglore hotel details of the distribution of percentage in terms of excellent, very good, average and poor of each individual hotels falling on the y-axis against the hotel names on x-axis.

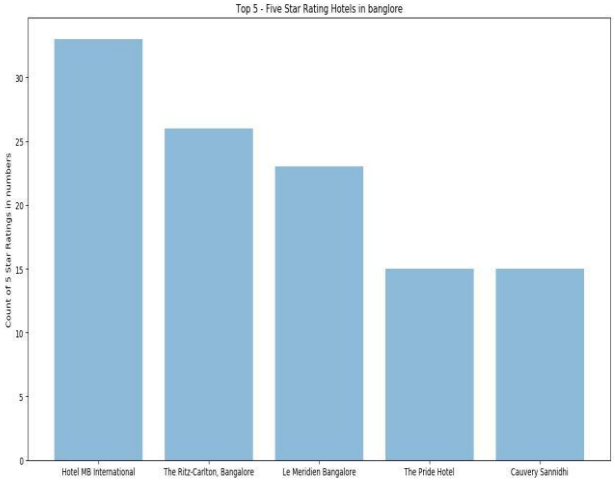

Fig.6 Top five 5-star rating hotels in Banglore

Fig.6 represents top 5-star ratings hotels in Banglore against the count of peoples who have give it. Graph gives the information about the top 5-star rating hotels in Banglore. As u can see in graph, y-axis defines the total number of people who has given the five star rating to particular hotels, and x-axis gives the names of the hotels to which maximum 5-star ratings is given.

The figures 7 and 8 shows the visualization of word frequency using one of the visual form that is Word Cloud. Here in our paper we took a single comment means a single review description of one of the user and applied word cloud technique, As followed below in the Fig.7 you can see the frequency of the word that is repeated at maximum times will be displayed bigger and better compared to the other words in the comment. The words that are repeated in the given input comment is 26th, 25th and stay etc.

```
Out[41]: 'I booked a room on 25th January for 26th stay directly with Taj Hotels. When I reached on 26th afternoon and gave my confirmat
         ion slip at reception, I was shocked to find that by mistake I had put the date of 25th for my stay....More'

In [43]: gen_wordcloud(comment) # this generate the required word cloud, as you can see Indian and food are most frequent words
```

Fig.7 Word Cloud representation for a single comment

In the second Fig.8 of the Word Cloud visualization we combined all the descriptions using we have collected from the website through join function and applied word cloud technique, there we experience some words are bigger in size and those are the repeated the most and smaller size words are at minimum repetition in the data The words that are repeated at most in the given input description is Hotel, room and good etc.

```
In [44]: gen_wordcloud(' '.join(pd.read_csv('individual_review_details.csv').Description))
```

Fig.15 Word Cloud representation for all the comments combined together

Sentimental Analysis is all about feeding the sentence to the sentimental analysis classifier, it does the sentimental analysis and classifies whether the given sentence is positive, negative or neutral. You can see in the below Fig 9 we have checked for one of single review comment given by the user on some particular hotel and Fig 10 we fed three comments in general to show all the possibilities of emotions like positive, negative, and neutral and classifier also gives us how confidently it is a particular emotion in number.

```
from vivekn_sentiment import ViveknSentimentAnalyzer

# This analyzer needs the files vivekn.sentiment.model.json

vivekn_analyzer = ViveknSentimentAnalyzer()

comment

'I booked a room on 25th January for 26th stay directly with Taj Hotels. When I reached on 26th afternoon and gave my confirmat
ion slip at reception, I was shocked to find that by mistake I had put the date of 25th for my stay....More'

vivekn_analyzer.analyze_sentiment(comment)

{'confidence': '70.2688',
 'sentiment': 'Negative',
 'text': 'I booked a room on 25th January for 26th stay directly with Taj Hotels. When I reached on 26th afternoon and gave my
confirmation slip at reception, I was shocked to find that by mistake I had put the date of 25th for my stay....More'}
```

Fig.9 Sentiment Analysis use case1

Here the revive that is given as an input is 'I booked a room on 25th January for 26th stay directly with Taj Hotels. When I reached on 26th afternoon and gave my confirmation slip at reception, I was shocked to find that by mistake I had put the date of 25th for my stay....More' and it is giving as negative emotion and with the confidence of '70.2688' percent out of 100.

```
vivekn_analyzer.analyze_sentiment('banglore hotels are very bad')

{'confidence': '90.7276',
 'sentiment': 'Negative',
 'text': 'banglore hotels are very bad'}

vivekn_analyzer.analyze_sentiment('i liked the garden area in one of the mysore hotel')

{'confidence': '89.3241',
 'sentiment': 'Positive',
 'text': 'i liked the garden area in one of the mysore hotel'}

vivekn_analyzer.analyze_sentiment('i will have food in banglore hotel ')

{'confidence': '51.0218',
 'sentiment': 'Neutral',
 'text': 'i will have food in banglore hotel '}
```

Fig.10 Sentiment Analysis use case2

In this section we have predicted the best hotels based on their star ratings .our last and final work is to show the users the best hotels in a particular district based on the review ratings given by the already experienced tourists or users in online tourism website. We have done this with the help of Logistic Regression and we have referred this algorithm that has been applied to banking sector in[8],

In our paper we have predicted the best hotels of 2 districts like Banglore and Madikeri (Kodagu). Among several hotels in Banglore district the top rated hotels are displayed in graph they are Shreyas Yoga Retreat, Fairfields by Marriott Bengaluru Rajajinagar, The Ritz Carlton Bangalore, Bengaluru Marriott Hotel Whitefield and so..on. Similarly for Madikeri the top rated hotels displayed in graph are Riyavar Luxury Homestay, Sneha Homestay, OYO 7170 Homestay Vintage Villa accordingly.

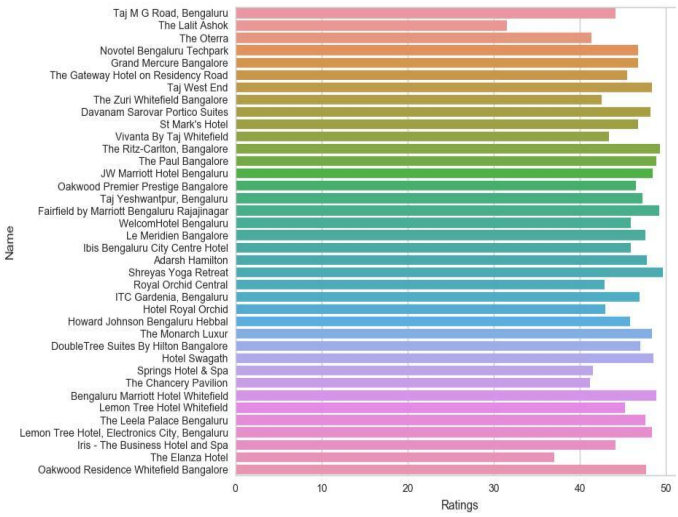

Fig.18 Prediction of best rating hotels in Bangalore on an average

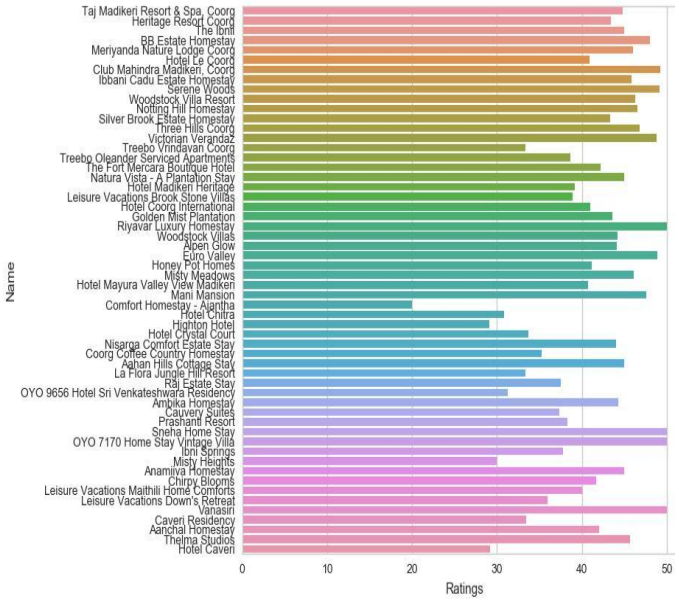

Fig.19 Prediction of best rating hotels in Madikeri on an average

V. CONCLUSION

Increasing in the development of the internet, there is an enormous data provided by the users in the form of reviews available on the websites of various fields. One among them is a tourism websites called Tripadvisor.in. When people has to decide which place is better to prefer is very much important things. Our research helps in quick analysis of the best hotels in different districts of Karnataka. The process started from extraction of reviews from the website and analyzing in different ways to suggest the best hotels among the available hotels. We applied different machine learning tools and algorithm to our research in order to produce better results.

The extracted data is analyzed in terms of Data Visualization in the form of Graphs, Word cloud, Sentimental Analysis and finally the Prediction of best hotels based on their star ratings.

REFERENCES

[1] Large Hotel Review Dataset [https://www.tripadvisor.in/Tourism-g297627-Karnataka-Vacations.html]

[2] [https://www.mysql.com/products/workbench/] https://dev.mysql.com/doc/]

[3] [http://nikgrozev.com/2015/12/27/pandas-in-jupyter-quickstart-and-useful-snippets/]

[4] [https://jakevdp.github.io/PythonDataScienceHandbook/04.00-introduction-to-matplotlib.html]

[5] Mohammad F. A. Bashri, Retno Kusumaningrum "Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization" International Conference on Information and Communication Technology (ICoICT), ISBN: 978-1-5090-4911-0 (c) 2017 IEEE, 2017.

[6] Deebha Mumtaza, Bindiya Ahujab "Sentiment Analysis of Movie Review Data Using Senti-Lexicon Algorithm" International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 978-1-5090-2399-8/16/$31.009 (c) 2016 IEEE, 2016.

[7] Vivek Narayanan, Ishan Arora, and Arjun Bhatia "Fast and Accurate Sentiment Classification Using an

Enhanced Naïve Bayes Model", International Conference on Intelligent Data Engineering and Automated Learning, H. Yin et al. (Eds.): IDEAL 2013, LNCS volume 8206, pp. 194-201, 2013. (c) Springer-Verlag Berlin Heidelberg 2013

[8] https://datascienceplus.com/building-a-logistic-regression-in-python-step-by-step Susan Li, Senior Data Scientist, Kognitiv Corp "Building a Logistic Regression In PYTHON, step by step Published on October 6, 2017.