Sentiment Analysis of Customers Opinions on Hotel Stays using Voted Classifier

Anurag Patel
Student,
Dept. of Information Technology,
Shah and Anchor Kutchhi Engineering college,
Mumbai, India.

Bhavya Chheda
Student,
Dept. of Information Technology,
Shah and Anchor Kutchhi Engineering college,
Mumbai, India.

Bhavik Jain
Student,
Dept. of Information Technology,
Shah and Anchor Kutchhi Engineering college,
Mumbai, India.

Manya Gidwani
Prof.,
Dept. of Information Technology,
Shah and Anchor Kutchhi Engineering college,
Mumbai, India.

Abstract—A trip always revolves around your hotel and selection for the same depends on factors like, distance from the place of a visit, quality, staff, rooms, etc. One of the most reliable and trusted ways to get information about a hotel is the opinion of people who have already visited that particular hotel. Opinions on booking websites are in the form of reviews that are sometimes short and sometimes very lengthy. Sentiment analysis of the reviews helps in understanding the reviewer's sentiment quickly. Therefore, accuracy of such a model should be as high as possible. The factors for selection are judged by the users on the basis of the sentiment being positive and negative which is well described in this paper. Further on, we elaborate the several algorithms that we used for sentiment analysis of hotel reviews, which minimizes the noisy data and classifies the reviews based on the model created. During the implementation, we trained six classifiers that gave an accuracy of around 88 - 94%. Further to reduce the margin of error and maintain highest possible accuracy along with it, the voted classifier was developed. The voted classifier included the top five best performing classifiers and gave an accuracy of 93.57%.

Keywords—Accuracy, classifier, hotel, opinions, reviews, sentiment, voted classifier

I. INTRODUCTION

The opinion of a user about a hotel or any place in the form of reviews is a widely trusted resource for business and analyzing these reviews is one of the most significant factors, reason being, 91% of people aged to their youth rely on online reviews. Out of which 46% just discard the option if its rating ranges lower than 4 stars out of 5. Making the factor tone of the comment, the most important reason being a negative tone, drives away the user psychologically and works vice versa for the positive review.

Business is the ultimate goal of any service and reviews have a dramatic effect on numbers associated with it. A positively scored review hikes your business whereas a negative review drives customers away. This fact proves that reviews give credibility to your business but does it mean that negative reviews are all bad? No, Not at all. In fact, the negative reviews bring in critical comments which can help an organization build on its weak spots. In the Hotel industry services and accessibility play a vital role as analysis of reviews helps to achieve more good and minimal bad. Analysis tools thus help in the summarization of an abundance of review data into brief statistics, helping any organization or user to conclude with the best decisions. Adding to which comes marketing, it can be a stronghold if you market with your positives to correct people.

Research has been put in for several years and the opinions of people have been recorded for the best services. But, this review system goes vague if the data is not processed and made to develop a proper rating, scoring or statistical visual. The sentiment analysis has two levels of sentiment from emotion and a calculated scoring for positive and negative statements. Tourism and bookings are one of the top-grossing industries [7] and unorganized data becomes irritating for users. The Traveloka search based on trip advisor states the use of a bunch of algorithms such as SVM and Naive Bayes [5] for feature extraction and sentiment analysis. The difficulty with the classification revolves around how well does the review tell its opinion and polarity? The classification is done first on the basis of what the opinion is and what does the opinion stand for, a positive or negative sentiment.

A hotel has customers daily and thus daily reviews, which produces a load of data thus making it practically impossible for a user to read and get the gist of the place in a jiffy. This is when sentimental analysis of hotel reviews pinches in. The segregation of a single review in the breakdown of adjectives, adverbs and verbs helps the machine understand the sentiment behind the review. Enabling an overall classified output of several objects taken into consideration by a user during the selection of a hotel.

There are a bulk of tools available in the market for analysis of data but a blind use of tools on irrelevant data doesn't give accuracy. A proper combination of tools with clean data can raise accuracy by a marginal amount. With 93.57% accuracy on classification of reviews by the use of Voted classifier, the objects are classified as positive and negative and are scored using additional libraries.

II. RELATED WORK

Training and modeling of data are the most stringent processes in analysis and several data scientists try to design best models for scoring and generalization of reviews for ease of the user. The main aspects of the research being covered in this block include similar kinds of developments on preprocessing, feature-based extraction, and classifiers. The file is first extracted from its original type to a .txt file. The data extracted is labeled into positive and negative [6] coming from its original source and processed accordingly .

The abundant data found on the internet is not suitable for text classification because it contains duplicate and noisy data that needs to be eliminated for the best analysis. The data is cleaned with redundancies and ambiguity [7]. The preprocessing of data by cleaning it and working on sentence based and aspect-based classification is discussed. The sentence elements are tokenized and POS tagged for breaking the sentence in adverbs and adjectives[5][9], eliminating stop words.

The distribution of positive and negative words plays a vital role in classification of scoring aspects and extracting words in feature sets [5][6]. The labeled negative and positive data makes the further process easier as the data is then formatted to obtain a feature opinion pair [9].

Feature extraction works on the frequencies of words being repeated in the corpus [5][6], and this feature set is further used in the training of the reviews which later comes handy for testing. Further research takes place by Naive Bayes and Decision trees with few if-else statements for special cases [6]. The final stage of the process consists of the classifiers extracting features from the reviews and finding them in feature sets. The models are trained and tested using several classification algorithms namely Naive Bayes, Decision tree, Logistic regression, Support Vector Machine[5][6][9]. These models once executed on the reviews classifies them into positive/negative and the accuracies of the respective models are obtained
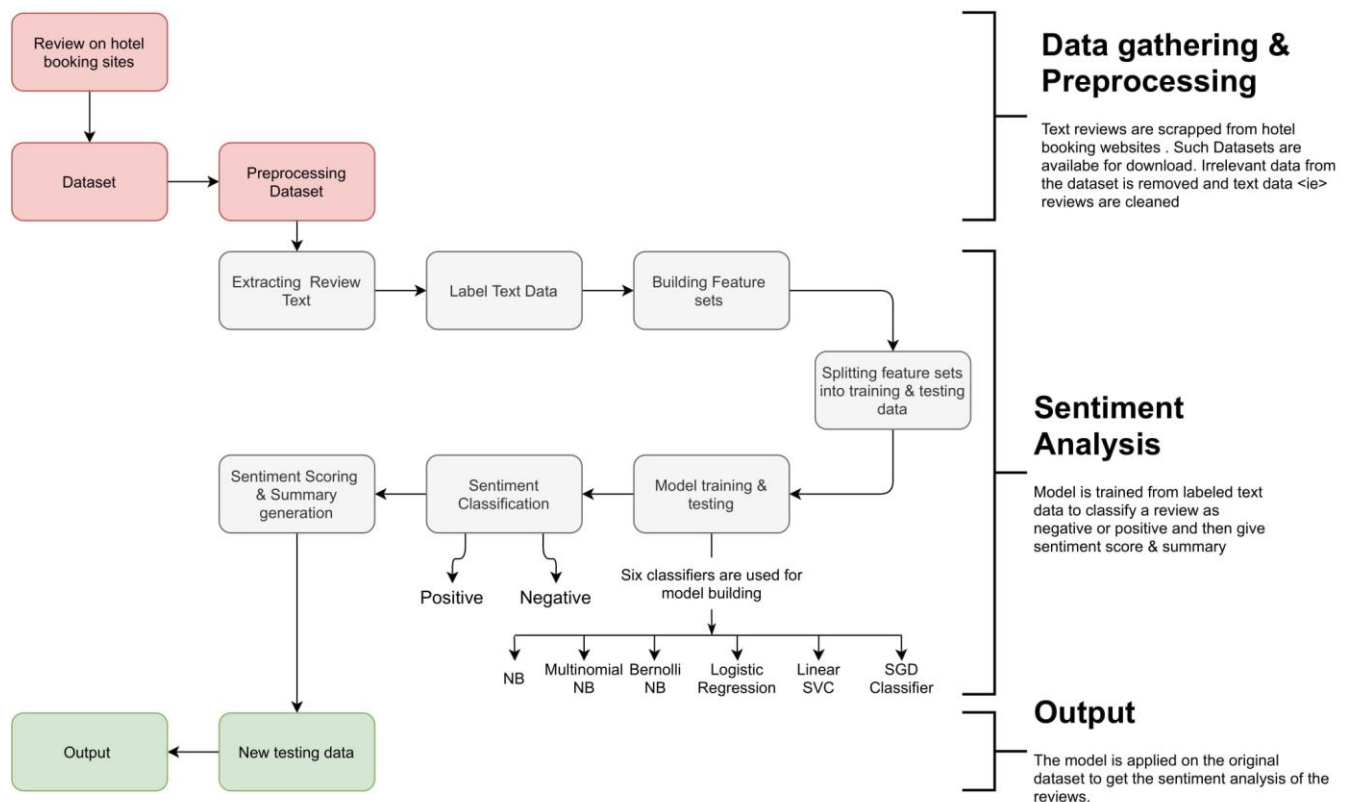
III. METHODOLOGY

Fig.1.System Architecture

A. Dataset

The dataset used for sentiment classification of hotel reviews is "515k hotel reviews Data in Europe" created by Jason Liu. The .csv file consists of 5.15 Lakh entries with a total of 17 columns namely Hotel Address, Review Date, Average Score, Hotel Name, Reviewer Nationality Negative Review, Review Total Negative Word Counts, Positive Review, Review Total Positive Word Counts, Reviewer Score, Total Number of Reviews Reviewer Has Given, Total Number of Reviews, Tags, days since review, Additional Number of Scoring, latitude, longitude. The dataset already has all the punctuations removed. This dataset was worked upon with a total of seven classifiers to achieve the best accuracy and minimum margin of error.
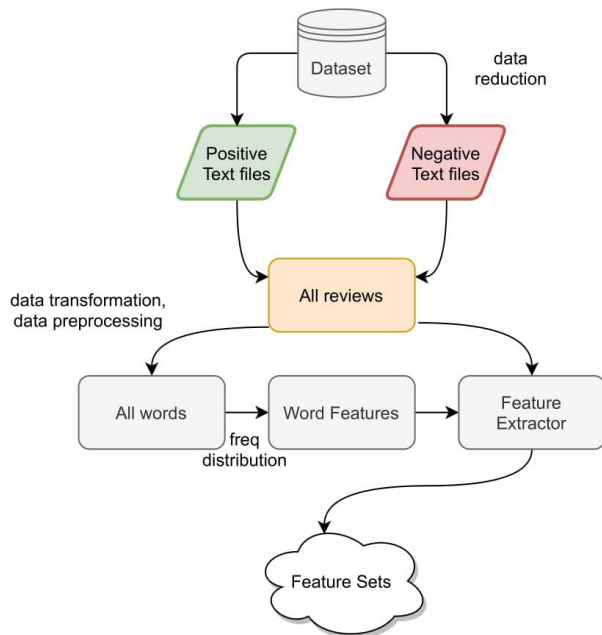
Fig.2.Data Preprocessing & Building Feature sets

B. *Data Preprocessing*

Data Preprocessing consists of 3 main stages i.e., Data cleaning, Data transformation, Data reduction. First step involves data reduction i.e., reducing the number of columns from the dataset and selecting 20000 random rows because the whole dataset is too large to be trained. It is necessary to reduce the columns because only the columns with name Positive Review and Negative Review contains the text on which the classification will occur, all other columns are helpful in understanding the data and for data analysis but they are not useful in sentiment classification. The data from positive reviews and negative reviews is extracted in two separate .txt file one containing the positive reviews while the other one contains negative reviews. While extraction reviews containing sentences like "No positive" / "No negative" were discarded. In the second step data transformation and data cleaning takes places simultaneously. The .txt file containing negative and positive reviews are opened in python, converted to lowercase, labelled with negative and positive label and appended in a single list in the list [review, label] format. This list is then used to extract all the words used in reviews both negative and positive. This task involves using word tokenizer, POS tagger, and then only appending adjectives, adverb and verb to a new list. These are achieved by the use of libraries named pandas and NLTK.

C. *Feature building*

For building the feature sets, frequency distribution is applied to the list containing all words. Once frequency distribution is applied the list is now sorted in an order where the words that occur the most are at the start of the list and least occurring words are at the end of the list. The top 5000 words from this list are then used for feature building. A function for finding features is created which accepts text as input parameter and returns as features only those words present in the top 5000 words. This function also contains data cleaning methods like removal of punctuation, converting the reviews in lowercase, word tokenization. This is important because when the model

will be used with unknown data then review entered can contain punctuation and that will hamper the feature extraction process. The next step is to build a feature set which will be used to train and test the model. This is done by applying find feature function to the list containing all reviews and their label. After the feature set is generated a random shuffle method is used on the feature set to avoid bias. The length of the feature set is 33658 i.e., the total of positive and negative reviews.

D. *Training and Testing of Data*

The feature set is divided into training set and testing set in the ratio 75:25 respectively. Instead of using the split function to divide the data, slice operator is used to select the first 25000 features of the feature set as training data. The remaining 8658 reviews are assigned to the testing set. Since the data has already been shuffled before assigning it to the testing set and training set, chances of bias have drastically been reduced. The training set is used to train a number of classifiers namely Naive Bayes, Bernoulli Naive Bayes, Multinomial Naive Bayes, Stochastic Gradient Descent, Logistic Regression, and Linear SVC and finally a Voted classifier which is a collection of other trained classifiers.
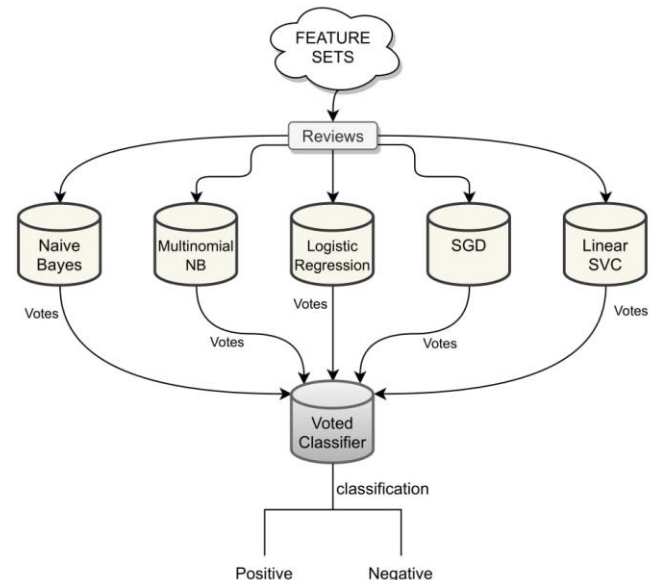

Fig.3.Voted classifier working model

E. **Model Construction**

1. *Naive Bayes:*

The Naïve Bayes classifier follows a simple assumption that each feature is independent of other features and it makes an equal contribution to the final classification or result. It finds the probability of an outcome when other probabilities are already given. Despite the assumptions made by the classifier are not valid in the real world but it has worked well in practice.

We have also used two variations of Naïve Bayes classifiers

> i. *Multinomial Naïve Bayes*
> In Multinomial Naïve Bayes classifier instead of features the feature vectors represent the frequencies with which events have occurred and represent it with

a multinomial distribution. Therefore, makes itself usable with features like word counts in text classification

ii. Bernoulli Naïve Bayes

In Bernoulli Naïve Bayes the features represent independent Boolean values and share a lot of features with the logistic regression technique. Establishing its similarity with the multinomial model this classifier uses binary term occurrence feature while the former uses term frequencies.

2. Logistic regression:

Logistic regression works as a supervised ML classification algorithm. In its basic form logistic function is used by logistic regression to model a binary dependent variable. Binary dependent variable means the target can only take two values either positive or negative. Positive classification by the model is a linear combination of one or more independent binary variables.

3. Stochastic gradient descent:

Stochastic gradient descent iterates with a larger number of iterations then compared to any other gradient descent. Though it takes a high number of iterations to reach the conclusion it is computationally less expensive and accurate because in every iteration the batch size is one. The path of SGD is also very noisy in comparison to other methods but in reality the path doesn't matter if the descent is able to reach the minima and that too in a shorter time period. It has advantages of: -

1. Efficiency.
2. Ease of Implementation.

4. Linear SVC:

Linear Support Vector Classifier is a Support Vector Machine that constructs a hyperplane in two-dimensional space for segregating feature into positive negative and gives an optimal hyperplane in an iterative manner used to reduce any error, where the goal is to select a hyperplane with the maximum possible margin among the support vectors in the feature set.

Linear SVC is a support vector classifier where the kernel = linear.

5. Voted Classifier:

Voted classifier is built by creating a custom class which takes as input the top five best performing classifiers i.e., Naïve Bayes classifier, Multinomial Naïve Bayes classifier, Logistic regression classifier, Linear SVC, Stochastic Gradient Descent classifier. These classifiers classify the reviews and their classification are votes to the voted classifier. Voted classifier classifies a review to be positive or negative based on maximum votes. The advantage of using the voted classifier is to increase the confidence in classification i.e., if a review is classified to be negative or positive by a voted classifier, it means 3 or more than 3 classifiers have classified it as negative or positive and hence that sentiment can be trusted more.

The following table shows the comparison of all the classifiers with respect to accuracy and runtime.

Table-I: Comparison of different classifier's Performance

Sr.no	Classifiers	Performance	
		Accuracy(%)	Run Time(s)
1	Naive Bayes	88.78	196.02
2	Multinomial Naive Bayes	89.29	102.45
3	Bernoulli Naive Bayes	88.72	105.10
4	Logistic Regression	93.39	119.11
5	Stochastic Gradient Descent	93.27	96.03
6	Linear Support Vector Classifier	93.97	183.70
7	Voted Classifier	93.57	190.16

6. Sentiment Scoring

Sentiment Score helps in understanding to what level the review is positive or negative. For getting the sentiment score TextBlob (A python package) has been used. TextBlob takes as an input a string and converts it to a blob. Blob.sentiment.polarity(text) method returns a sentiment polarity between -1 to +1, where -1 indicates very negative sentiment and +1 indicates very positive sentiment. Using mathematical operators of addition, multiplication and division score can be derived from the sentiment polarity with the range of 0-10.

F. Graphical User Interface

A GUI is made to test the model with various unknown testing data. GUI is made using inbuilt python libraries. The GUI contains features like a textbox to enter multiline review and buttons like "New", "Submit", "Clear", "Exit". As the name suggests the new button is used to write a new review, the submit button on click performs sentiment analysis of the reviews, the clear button clears the text in the textbox and the exit button terminates the program and exits the application. The GUI is named as the sentiment analyser. It shows as output three important factors to describe the user sentiment. The GUI imports the trained model and the functions to get a sentiment score and the summary of the review. On clicking the submit button the text from the textArea is passed as parameter to various sentiment analysis functions imported from the model. Features are extracted from the text and our classifiers classify the sentiment and give their vote to the voted classifier. Voted classifier returns the chosen sentiment, while sentiment score and sentiment summary functions return their respective analysis. All this is displayed below the button in the GUI.

The designed GUI can further be converted into an executable program once the models are exported as packages and available for download from the python packages repositories.
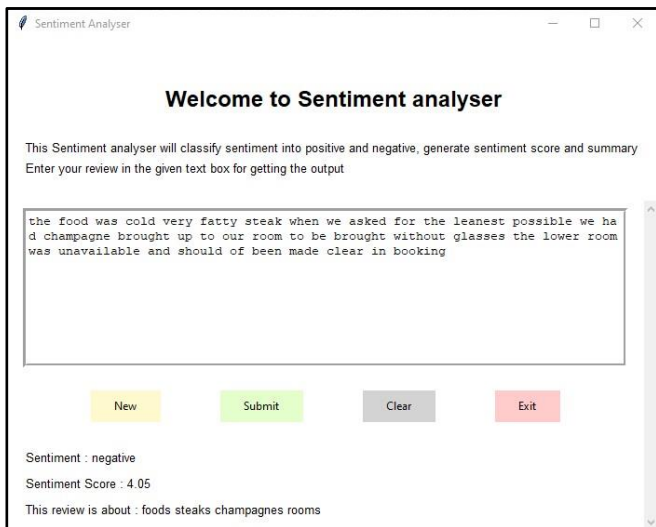
Fig.4.GUI Snapshot

IV. EXPERIMENT AND RESULTS

The dataset contains 515k review of luxurious European hotels but the total number of reviews extracted from the dataset for training and testing the model is approximately 40000 i.e., 20000 negative reviews plus the 20000 positive reviews, but after removing stand-alone sentences like "No negative", "No positive" from negative and positive reviews column in the dataset only 33658 reviews are remaining. From these reviews a list of all words is created but this list initially had only adjectives but later to increase accuracy adverbs were added and finally verbs were also added in the list for better accuracy and rich feature set. The list containing all words was subjected to frequency distribution to arrange the words in descending order of their occurrence. Word feature list was created by selecting the top 5000 words from the all words list after frequency distribution and these word features were used to create the feature sets. Above discussed cases and their defects are as follows.

A. Case 1

In Case 1 the allowed word type are adjectives. Adjectives are an important part of speech in understanding the sentiment of the sentence as they describe the customer's feelings towards the hotel. But testing the models trained only on adjectives with unknown data didn't reflect their high accuracy ranging from 87% to 92%. On further investigation, it was found that many sentences lacked the amount of adjectives needed for a proper classification. This was observed particularly in short reviews.

Table-II: Case 1 Comparison of classification by seven classifiers vs real values (positive:1, negative:0)

Sr.no	Real values	Classifiers						
		NB	MNB	BNB	LR	SGD	Linear SVC	Voted
1	1	1	1	1	1	1	1	1
2	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1
4	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1
9	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0

B. Case 2

In Case 2 the allowed word types included adjectives and adverbs both. In addition to adjectives, adverbs were also added in allowed word type as they modify the meaning of adjectives in many sentences and the same was reflected while training and testing the model. The accuracy of classifying the sentiment was increased marginally, but testing these models with unknown data did show an improvement over the model trained only on adjectives but still lacked in understanding the context of sentence and also the problem of lack of features to classify sentiment correctly persisted in short reviews.

Table-III: Case 2 Comparison of classification by seven classifiers vs real values (positive:1, negative:0)

Sr.no	Real values	Classifiers						
		NB	MNB	BNB	LR	SGD	Linear SVC	Voted
1	0	1	1	1	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1
5	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1
9	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1

C. Case 3

In Case 3 the allowed word types included adjectives, adverbs and verbs. Adding verbs to the allowed word types didn't give a significant increase in accuracy but it was still an improvement from case 2. The real difference was seen when these models were tested with unknown data. With verbs added in the allowed word types the feature set now had verbs in it which in turn made it possible to classify shorter reviews. This happens because even though the number of adjectives and adverbs remains low but the verbs in sentences allow the models to classify real world data satisfactorily.

Table-IV : Case 3 Comparison of classification by seven classifiers vs real values (positive:1, negative:0)

Sr.no	Real values	Classifiers						
		NB	MNB	BNB	LR	SGD	Linear SVC	Voted
1	0	1	1	1	0	0	0	0
2	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1
8	0	1	1	1	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0

Now, comparing the accuracies of the seven classifiers in different cases to get a perspective of improvement in the model's classifications over the three versions.

Table-V: Accuracies of classifiers in different cases

Sr.no	Classifiers	Accuracy		
		Case 1	Case 2	Case 3
1	Naive Bayes	90.44	89.31	88.78
2	Multinomial Naive Bayes	87.49	89.40	89.29
3	Bernoulli Naive Bayes	90.29	89.20	88.72
4	Logistic Regression	90.89	92.96	93.39
5	Stochastic Gradient Descent	90.79	92.73	93.27
6	Linear Support Vector Classifier	91.73	93.60	93.97
7	Voted Classifier	91.38	93.33	93.57

D. **Result**

In the comparison of classification by classifiers in different cases (Table-II, Table-III, Table-IV) it is observed that the voted classifier performs best in case 2 and case 3. The voted classifier is expected to reduce the margin of error by considering the classification of top performing classifiers and taking votes from them i.e., if it's positive or negative. If the number of votes for positive classification is more than the number of votes for negative classification then the voted classifier classifies it as positive and if it is the other way round then the voted classifier will classify it as negative. This is clearly observed in Table-IV Sr.no '1' and '8'. We have 6 classifiers, so there is a chance that 3 of them can classify a sentence as positive and the other 3 can classify it as negative. So to avoid this situation only the top 5 performing classifiers are allowed to give the vote.

sentiment	sentiment_score	pos_summary	neg_summary
positive	6.32	['rooms', 'staffs', 'locations', 'c	['bites', 'breakfasts', 'costs']
negative	3.5	['staffs', 'areas', 'hotels']	['rooms', 'bacons', 'breakfasts']
negative	6.16	['distances', 'walks']	['services', 'wasns', 'stars', 'trans
positive	6.75	[]	['Carpets', 'rooms', 'attentions']
positive	6.2	['staffs', 'carparks', 'goods']	['suites']
positive	5	['Locations', 'spots']	[]
positive	9.78	['Everythings', 'Locations']	[]
negative	7.75	['kinds', 'staffs']	['windows']
negative	5.86	['breakfasts', 'staffs', 'lies', 'bef	['windows', 'airs']
negative	2.95	['lounges']	['Costs', 'rooms', 'sizes', 'airs', 'c

Fig.5. Snapshot of newly added column

The final model along with the scoring function and summarizing functions are applied to the original dataset reduced to a sample of 1 lakh rows (to save computational cost and time). While applying the model and supporting functions to the generated sample of original dataset 4 additional columns are added in the process namely, "sentiment", "sentiment score", "positive summary", "negative summary". Sentiment column contains sentiment classification of the reviews. The Sentiment score column contains the score of the reviews. The positive summary and negative summary column contains those words that help in understanding in short what the review is about.

Table-VI: Sentiment analysis of testing data in different cases

Sr.no	Customer Reviews	Sentiment Analysis			
		Sentiment Classification			Sentiment Score
		Case 1	Case 2	Case 3	
1	Cost for room size shocking noisy air con in rooms Exec lounge is a waste of time breakfast was cold and no range of foods The sky lounge but highly overpriced	negative	negative	negative	2.95
2	Wish the pillows have been thicker The location is great	positive	positive	positive	9.0
3	Location friendliness and attentiveness of staff great breakfast loved the bar and the barman was exceptionally friendly	positive	positive	positive	8.12

4	good location and excellent staff	positive	positive	positive	9.25
5	Parking facilities are bad so Using a car to get to the hotel is a bad idea The experience at hotel was pleasant	positive	negative	negative	1.5

V. CONCLUSION

The Voted classifier from the 3rd model was successful in achieving the highest possible accuracy with minimum margin of error as seen in Table-V. Experiments were conducted on all versions of models with testing data and unknown testing data (real world data). The results from the experiments were used to improve not only the accuracy of models but also the quality of the feature sets. Further to make the analysis more informative, sentiment score and summaries were also added to the dataset as well as in the GUI using additional libraries. Although basic but the final model fulfills the task of performing sentiment analysis on large amounts of customer reviews with accuracies as high as 93.57%. This model is open for further development and the future scope includes binary sentiment classification at the aspect level which will throw new challenges like aspect identification and classification. It will also be focused on maintaining or improving the accuracies, optimizing the code and speeding up the process of building the model. The later stages of development will also include the creation of a library for easy installation and usage of the model for the sentiment classification & sentiment scoring of reviews given by the user.

This paper proposed a framework to perform sentiment analysis on customer's opinion on hotel stays. For creating a model for this framework a dataset with 5 lakhs plus entries has been used. In this framework, after loading the dataset and extracting the corpus for both negative and positive reviews, basic preprocessing steps such as conversion of words to lowercase, removal of punctuations, word tokenization, POS tagging were performed. After preprocessing, creation of feature sets with high-frequency features was achieved. These feature sets were shuffled using random methods to avoid any kind of bias and they were divided into training and testing sets for model training. Training various classifiers like the Naive Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Linear SVC, Stochastic Gradient descent and using them to build Voted classifier enabled testing and process for achieving high accuracies and trusted classification in the domain of sentiment analysis.

REFERENCES

[1] Mohey El-Din, Doaa. (2016). A Survey on Sentiment Analysis Challenges. Journal of King Saud University - Engineering Sciences. -. 10.1016.

[2] Walaa Medhat , Ahmed Hassan, Hoda Korasey, Sentiment analysis algorithms and applications: A survey , *Ain Shams Engineering Journal Volume 5, Issue 4*, December 2014, Pages 1093-1113

[3] V. Agarwal, P. Aher and V. Sawant, "Automated Aspect Extraction and Aspect Oriented Sentiment Analysis on Hotel Review Datasets," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-4.

[4] B. Devkota, K. Kim, C. Zhuang and H. Miyazaki, "Disaggregate Hotel Evaluation by Using Diverse Aspects from User Reviews," *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Kyoto, Japan, 2019, pp. 1-8.

[5] S. Kurniawan, R. Kusumaningrum and M. E. Timu, "Hierarchical Sentence Sentiment Analysis Of Hotel Reviews Using The Naïve Bayes Classifier," *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2018, pp. 1-5

[6] W. Songpan, "The analysis and prediction of customer review rating using opinion mining," *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, London, 2017, pp. 71-77.

[7] M. Afzaal, M. Usman and A. Fong, "Tourism Mobile App With Aspect-Based Sentiment Classification Framework for Tourist Reviews," in *IEEE Transactions on Consumer Electronics*, vol. 65, no. 2, pp. 233-242, May 2019.

[8] A. Nazir, Y. Rao, L. Wu and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," in *IEEE Transactions on Affective Computing*.

[9] N. H. A. Rahim and M. A. F. Hasnan, "A Summarisation Tool for Hotel Reviews," *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*, Guangzhou, China, 2018, pp. 218-223.