Sentiment Analysis of Restaurant Review

Divyanshu Shukla  
Computer Science  
Noida Institute of Engineering and TechnologyGreater Noida,Uttar Pradesh, India  
divyanshu2003mth@gmail.com

Abhijeet Kumar  
Computer Science  
Noida Institute of Engineering and TechnologyGreater Noida, Uttar Pradesh, India  
ak9111709@gmail.com

Aayush Solanki  
Computer Science  
Noida Institute of Engineering and TechnologyGreater Noida, Uttar Pradesh, India  
aayushsolanki.sa@gmail.com

*Abstract*— In this paper we did sentiment analysis of restaurant review. We all visit restaurant with our family and friends to eat something. But before visiting a restaurant we choose which restaurant we have to visit. So, we choose restaurant on the basis of reviews made by customers who have visited that restaurant earlier. Restaurant is a businesses so restaurant owner allows customers to review about their experience related to restaurant so that they can make changes according to that in their menu and facilities. So to classify which review is positive or negative and would have positive or negative impact on others they use sentiment analysis. Sentiment Analysis is way to analyze which review is positive or negative using natural language processing and machine learning algorithm. In this research paper we have used Naïve Bayes algorithm(GaussianNB and MultinomialNB), Support vector classifier from Support vector machine, Kneighbours Classifier and Logistic Regression and this we have taken dataset from Kaggle.

Keywords—Support vector machine, Naïve Bayes, Sentiment Analysis, Natural Language Processing

# Introduction

Customer expectation and satisfaction is much important for restaurant owners. In today’s generation customer write their opinion on in the form of review. For writing review customers uses online media such as twitter or on google which increase popularity of the product and services sell by the restaurant owner.

Restaurant is a basically a businesses that serves food for customers and exchange a certain amount of money from them. Although improving quality form review is very useful for restaurants but only very few restaurant use this method to improve their services. Also, many algorithm might be used for this study.

We have collected data from Kaggle and remove did data cleaning using preprocessing such as remove stopwords, remove punctuation, remove fullstop etc. using Natural language Toolkit which is Python library. After data cleaning we used machine learning algorithm such as Naïve Bayes(GaussianNB and MultinomialNB), Support Vector Calssifier from Support Vector Machine, Logistic Regression and KneighboursClassiefier. All these algorithm are imported from with help of sklearn which is Python library used for Machine Learning.

The main aim of this research paper is to analyze best method for restaurant review using Machine Learning algorithm.

# Literature Survey

Customer satisfaction is essential in the field of marketing. The satisfaction of customer can be done by knowing the demand of customers based on review written by them about the services provided by restaurant.

Data analysis of restaurant review is done by using sentiment analysis to know the satisfaction and the demand of customers. By doing data analysis restaurants can increase their profits by providing services demanded by customers. Data analysis can be done by sentiment analysis technique

Sentiment analysis basically used to find the positive or negative polarity of the sentences(or Review) by removing or omitting unnecessary part of the sentences using Natural Language Processing(NLP) or some Machine Learning algorithm

In past sentiment analysis is used for movie reviews to predict the response of the movies as positive or negative. Various Lexicon based approaches are used to review the comments of the movie and classify them as positive or negative. Lexicon based approach like VADER, Textblob and NLTK are tested for this purpose and VADER gives best accuracy [1].

Sentiment analysis technique is used for the analysis of comments received on a YouTube video, the reviews are done not only for English language comments but also for the other languages like Indonesian, etc. This analysis is conducted using various techniques and results shows that most of the approaches in literature are under simple sentiment analysis [2]. This paper also suggests to review comments of some other different languages like Arabic

One of the useful areas of sentiment analysis is in the field of medical science. Sentiment analysis can be used to predict the drug for any type of illness. This is done by reviewing the public comments and reviews regarding any drug used by them during any illness and the effectiveness of that drug to cure that illness. This can be a very effective way of curing illness if this technique is used in mass [3].

Depression now-a-days are very common and serious problem specially for the young population. One problem with depression is that we can’t predict whether a person is in depression or not, but with the help of machine learning techniques we can examine a person social media post and predict whether he is depressed or not. This is very much possible as we can see social medias are new diaries these days where a person expresses his feeling. By carefully examining a person social media with the help of machine learning algorithms we can predict whether he/she is depressed or not [4].

Sentiment analysis is largely used during Covid-19 pandemic to predict the feelings of people regarding this deadly pandemic and it’s effect on economic, social, political and technological levels. Data collected from twitter and other sources are predicated using various machine learning techniques as positive and negative. Analysis was done between January and July and result shows that only January is the month in which negative thoughts predominated beside January all other months saw a rise in positive thoughts, this positive thoughts shows that people were spreading hope in tough times [5].

Sentiment analysis is also used to predict people perspective regarding different types of Covid vaccines developed by pharma companies like Pfizer, Moderna, and AstraZeneca. Tweets of people were analysed to predict the safety and effectiveness of the vaccines, this was mainly done to know about the best vaccines by analysing tweets of people regarding these vaccines using various machine learning techniques [6]. Results predict that Pfizer gets more positive reviews while AstraZeneca gets least.

Sentiment analysis techniques are used to predict the people opinions regarding the release of new iPhone as positive and negative. This can be done by anlysing various tweets of people and other reviews of new iPhones on social media by applying various machine learning techniques.This analysis help other people whether they have to buy new iPhone or not [7].

Emotion detection is also done using various machine learning techniques. Emotion of a person can be detect using facial expressions, hand momens and by analyzing his day to day behaviour. The emotion detection market is growing very fastly now a days, various healthcare companies are using these emotion detection methods to sold their products to the required person. Doctors are using these technique to predict about the problem or diseases a person is suffering from and give proper treatment according to the requirements [8].

Machine learning and sentiment analysis techniques were used to see the impact of President Trump’s tweets effect on US stock market. It is analyzed that tweet has a negative impact when it is tweeted by President during open market hours while tweets with strong positive or strong negative sentiment has a positive impact on the stock market [9].This is used to predict the effects on few companies by collecting President Trump tweets between some particular period of time.

Sentiment analysis plays an important role during Covid-19 to predict whether a person is suffering from anixety or not. This is done mainly by analysing his social media persence, we analyse his tweets and comments to predict the anxiety.This plays a key role in government success of a program during Covid, government tries to figure out the anxiety of a person and tries to improve it by giving him the better tretment. If a person is well by health, he will welcome and benifited well by government programs [10].

# methodology

In this paper we have used many algorithm such as Naïve Bayes, LogisticRegression, Kneighbours Classifier and SVC.There are several methods for Research methods

## DATA COLLECTION

We have collected datasets from Kaggle which is named as Restaurant\_Reviews.tsv. It has 1000 reviews. On Kaggle Logistic Regression and Multinomial Naïve Bayes algorithm were already used.

## DATA PREPROCESSING AND ANALYSIS

On the preprocessing there are several steps that everyone must follow while preprocessing data for sentiment analysis. First we have removed stopwords which have not any value in review. We remove all those values using natural language toolkit(nltk) which is python module and provide stopword in for different languages. We did stemming and lemmatization on words. Stemming and Lemmatization are different from each other. Stemming removes the last few characters like “es” and “-ing” which leads to incorrect and spelling error. Lemmatization understands the word and convert it in to their root form like it converts “Loved” into “Love”. Example given below for data cleaning: “Wow….Loved this place.” became “Wow love place” It remove dots, full stop, this and change Loved into love.

The remove punctuation is a process of removing of punctuation that appear often many times and usually have no meaning like “- , … , /”.

## SENTIMENT ANALYSIS

The cleaned dataset have given number 0 and 1 for polarity where 0 means negative polarity and 1 means positive polarity. On the basis of these polarity in train dataset the polarity of test dataset and fresh data had determined from classification technique with help of different algorithm

## CLASSIFICATION

#### Naïve Bayes: It is a supervised learning approach for classification. It is based on the Bayes theorem. It is the most straightforward and efficient classification technique, which aids in the development of quick machine learning algorithms for predictions. Text classification is its principal use. It is a probabilistic classifier, which means that it makes predictions based on the likelihood that an object exists. The best application of Nave Bayes is sentiment analysis.

It is known as the Nave Bayes theorem because it combines the words "Nave" and "Bayes," which mean:

Naive: It is referred to as naive since it makes the assumption that the existence of one feature is independent of the occurrence of another.

Bayes: It is known as Bayes since it utilises the Bayes theorem's basic tenet.

It is possible to calculate the probability based on prior knowledge using the Bayes Theorem, sometimes referred to as the Bayes Rule or Bayes Law. Conditional probabilities play a role.

P(A|B) is posterior probability: Probability of A when B event has occurred.

P(B|A) is Likelihood probability: Probability of B when A event has occurred.

P(A) is prior probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal probability: Probability of evidence.

We have used Naïve Bayes algorithm because it is fast and easy of to implement algorithm to predict a class of dataset. It is popular for text classification.

The Nave Bayes algorithm has the drawback of learning no association between the features because it presumes that all features are independent or unconnected.

#### Logistic Regression: It is a commnly used machine learning algorithm. It is a supervised machine learning technique. We have used this algorithm because it predict output of categorical discrete value. It means output must be discrete or categorical. It’s output must be Yes or No, 0 or 1, or True or False.

Logistic Regression is similar to Linear Regression but logistic regression uses more sophisticated function known as sigmoid function. Sigmoid function squeeze any value between 0 to 1. The Sigmoid function can be given as:

Graph of this function can be given as:

Figure : Sigmoid Function

There are three types of logistic regression. These are follows:

1. Binary Logistic regression: If there is outcomes like Yes or No(means Two or more binary outcomes), then logistic regression is called Binary Logistic Regression.
2. Multinomial Logistic Regression: If there is outcomes like first, second and third class or no class degree(means Three or more outcomes), then logistic regression is called Multinomial Logistic Regression.
3. Ordinal Logistic Regression: If there is outcomes like Multinomial Logistic Regression but here with the order like customer rating in super market(means Three or more outcomes), then logistic regression is called Ordinal Logistic Regression.

We have used Binary Logistic Regression for our project.

This model can work for all datasets , but still there are assumptions or requirement we have to consider for it to work well. Requirements for Logistic Regression to work well are as follows:

1. Binary Logistic Regression should have binary dependent variable.
2. There should only be relevant variables included.
3. The independent variable must be unrelated to one another.
4. The log chances are proportional to independent variables,
5. Logistic regression requires a large sample size..

It is simple to use, understand, and train. It is highly quick in categorising unidentified records and offers information on both the appropriateness of a predictor (coefficient size) and the direction of relationship (positive or negative). Although it can overfit in high-dimensional datasets, logistic regression has a lower tendency to do so. It performs well when the dataset is linearly separable and has good accuracy for many simple datasets. To prevent over-fitting these cases, one can think about Regularization (L1 and L2) strategies.

Logistic regression will not be used if the number of observations is less than the number of features since it may cause  over-fitting. Only discrete functions can be predicted with it. The dependant variable in a logistic regression is therefore restricted to a set of discrete numbers. Logistic regression has a linear decision surface, hence it cannot be used to address non-linear issues.

#### KneighboursClassifier: It is one of the leading ,flexible and straightforward machine learnig algorithm. It is used in variety of applications such as, healtcare, handwriting detection, image recognition etc. It is Non-parametric means the structure of model is determined from the dataset. It is Lazy algorithm too,means if we are generating model then we don’t need any training data points because all training data used in testing phase, because of this training phase become faster and testing phase become slower and costilier(means it take more time and memory).

In KNN, K is the number of nearest neighbors. It is the main deciding factor. Usually K is an odd number. It is known as K nearest neighbour, when value of K become one(1) i.e. K =1.

Suppose label will predict a point A. First, you have to find k points nearest to the point A and the points will be classified by majority votes of its Kneighbours. Each objects votes for their class and class with majority voting taken as prediction. To find closest similar points, we have to find distance between the points that can be find using distance formula.

Following steps are followed in KNN:

1. Calculating the distance
2. Finding closest neighbors
3. Voting for labels



Figure : KnearestNeighbours

#### Support Vector Machine: Support Vector Machine is the one of the most popular supervised algorithm. It is used for classification and Regression problems. We have used it for classification problem because we have to classify wheather the output is positive(0) or negative(1).

The main aim of the support vector machine is to create a best fit line so that new data point can be easily put in correct category in the future. The best fit line is known as hyperplane.

In SVM, hyperplane is made by extreme points that are chosen by support vector machine and these extreme cases is called support vectors and hence the algorithm is known as support vector machine.



Figure : Support Vector Machine

There are two types of SVM:

1. Linear SVM: It is used for dataset which are classified into two classes by using a single straight line, these dataset is is known as linear separable data and known as linear SVM classifier.
2. Non-linear SVM: It is used for dataset that cannot be classified by using straight line, these dataset is known as non-linear separable data and known as non-linear SVM classifier.

When there is clear margin of separation between the classes then Support Vector Machine works well. In high dimensional space it is more efficient and comparatively memory efficient.

Support vector machine algorithm is not suited for large datasets. It does not work well when dataset has noise.

# Result

The various machine learning algorithms like Gaussian NB, Support Vector Classifier (SVC) From Support Vector Machine, KNeighbours Classifier, Multinomial NB and Logistic Regression(L) gives different accuracy, shown in below chart (in %).

Figure 4: Accuracy Comparison of Different Algorithm

Evaluation metrics used here are confusion matrix, accuracy, precision and recall:

Using Gausian Naïve Bayes:

* Confusion matrix of prediction is :

[[55, 42], [12, 91]]

* Accuracy of prediction is 73%.
* Precision of prediction is 0.68.
* Recall of prediction is 0.88.

Using Support Vector Machine:

* Confusion matrix of prediction is :

[[89, 8], [36, 67]]

* Accuracy of prediction is 78%.
* Precision of prediction is 0.89.
* Recall of prediction is 0.65.

Using KNeighboursClassifier:

* Confusion matrix of prediction is :

[[77, 20], [42, 61]]

* Accuracy of prediction is 69%.
* Precision of prediction is 0.75.
* Recall of prediction is 0.59.

Using Multinomial Naïve Bayes:

* Confusion matrix of prediction is :

[[74, 23], [22, 81]]

* Accuracy of prediction is 77.5%.
* Precision of prediction is 0.78.
* Recall of prediction is 0.79.

Using Logistic Regression:

* Confusion matrix of prediction is :

[[80, 17], [29, 74]]

* Accuracy of prediction is 77%.
* Precision of prediction is 0.81.
* Recall of prediction is 0.72.

Support vector classifier from support vector machine gives best accuracy of 78% while KNeighbours gives least accuracy of 69%.The results obtained from any of the machine learning algorithms are analysed (for best accuracy one can prefer Support Vector Classifier(SVC) ) and we can easily predict whether the overall review of the restaurant is positive or negative. The overall precision obtained in this experiment is 0.89. Keeping in mind the accuracy of 78% and precision of 0.89, the Support Vector Machine model can be considered as a good model which fits for this type of analysis.

# Conclusion

The overall conclusion of this research paper is that after using the designed model one can easily predict the reviews as positive and negative. After predicting, we need to analyse the positive and negative parts of the restaurants and if we try to work on negative parts and improves it then it is going to be a phenomenal rise in the restaurants sales. This model has a great scope in future also, it can be used to predict the review of any services as positive and negative. It can also be used at government level to review the schemes and various services of the government and if it is not satisfactory then government can try to improve it.

# References

|  |  |
| --- | --- |
| [1] | V. Bonta, N. Kumaresh and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for," *Asian Journal of Computer Science and Technology,* pp. 1-6, 2019. |
| [2] | S. Garg, "Drug Recommendation System based on Sentimen Analysis of Drug Reviews using Machine Learning," in *IEE*, 2021. |
| [3] | R. F. Alhujaili and M. W. Yafooz, "Sentiment Analysis for Youtube Videos with User Comments : Review," in *IEEE*, 2021. |
| [4] | K. A. Govindasamy and N. Palanichamy, "Depression Detection Using Machine Learning Techniques on Twitter Data," in *IEEE*, 2021. |
| [5] | C. R. Machuca, C. Gallardo and R. M. Toasa, "Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach," *Journal of Physics,* pp. 1-7, 2020. |
| [6] | F. M. J. M. Shamrat, S. Chakraborty, M. M. I. J. N. Muna, M. M. Billah, P. Das and M. O. Rahman, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science,* pp. 463-470, 2021. |
| [7] | W. Bourequat and H. Mourad, "Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine," *International Journal of Advances in Data and Information Systems,* vol. 2, pp. 36-44, 2021. |
| [8] | K. G, P. K. K, S. J and S. P, "Emotion and Sentiment Analysis using Machine Learning," in *IEEE*, 2021. |
| [9] | J. D. Kinyua, C. Mutigwe , D. J. Cushing and M. Poggi, "An analysis of the impact of President Trump’s tweets on the DJIA and S&P 500 using machine learning and sentiment analysis," *Journal of Behavioral and Experimental Finance,* pp. 1-14, 2021. |
| [10] | S. Saifullah, Y. Fauziah and S. A. Aribowo, "COMPARISON OF MACHINE LEARNING FOR SENTIMENT ANALYSIS IN DETECTING ANXIETY BASED ON SOCIAL MEDIA DATA," *Jurnal Informatika,* vol. 15, pp. 45-55, 2021. |