

Group Project on

Natural Language Processing (CS 491)

Academic Year- 2020-21

On

**Restaurant Review Prediction**

Submitted By:

**Group No**. - 18

Shrey Bhardwaj (BT18GCS092)

Aniket Chakraborty (BT18GCS081)

Swattik Chakrabarty (BT18GCS085)

**Introduction**

In today’s digital era of electronic word of mouth rather than going for recommendations as ewom have outweighed the traditional marketing strategies regarding influence on customer’s purchase decisions.

According to the trip advisor’s marketing survey in 2017 it is said that, based on online review, 94% of the diners choose a restaurant. So we can only imagine what is the condition now. Restaurants and customers according to multiple surveys conducted also agreed that online listing service is one of the most effective marketing channels for businesses as it drives food business to the local restaurants who have lost crowds to the big restaurant chains/franchises like KFC and Dominos.

In this era of everything available at the tip of our fingertips and having huge constraints of time, people neither have patience to sit and read reviews therefore it will be more feasible if that review can be converted into stars so that people can just see the rating and decide.

**Problem Statement**

* It is hassling and laborious to manually see millions/ billions of reviews online and categorize each of them into positive and negative
* According to Harvard Business School people who check online reviews before visiting a restaurant do not check the written reviews but directly see star rating and hence it is more beneficial for the owner of restaurant and customers to directly convert online reviews to numerical values and eventually to stars

**Literature Survey**

A few decades ago, researchers tended to conduct content analysis manually to identify the product or service features most important to customers based on the word frequency. To better understand aspects that contribute to a helpful review, machine learning with the help of NLP, allows a machine to extract and classify online reviews, has been utilized to provide more insights and make predictions from large amounts of reviews collected from customers. Compared to traditional forms of manual content analysis, machine learning methods for text data are less time consuming and labor intensive.At present in the era of internet eWOM becomes a popular mode of communication it's work is to defined positive or negative statement about the product or food based on customers review and at last gives us the percentage that how many customers are satisfied with this product or food. They give the percentage on the basis of algorithms in machine learning to predict the reviews of the data and researchers are still doing their research in this field to tell us which algorithm has the highest accuracy.

**Proposed Methodology**

**Data collection**

The data was collected from kaggle. In that data set there are 1000 reviews given by customers for a particular restaurant and each review has been classified into the class whether it’s positive or negative by assigning them a binary value of 1 and 0 respectively. After getting these reviews we plot a graph to get a better grasp of the data and after plotting we concluded that the data which we collected is uniformly distributed.

**Data Preprocessing**

With the help of regular expression in python all the characters except a-z and A-Z are removed from the entire dataset. All the characters are then changed into lower cases of their respective forms. Then they are split on the basis of space and stored in a list.

After that with the help of porter stemmer and the stopwords collection in nltk all the stop words are removed as they do not contribute to positive or negative sentiment of a review. The remaining words are changed to their root form and appended in an empty list called corpus.

An object of count vectorizer is created to make a matrix of the most frequent 600 words corresponding to each review in the dataset with the help of the corpus list. This matrix is filled with binary values to specify if that particular word is present in the review or not. This is called a bag of words and will be ultimately used to train our model.

**Classifier Set Up**

The classifier which we used is Gaussian naive bayes.Before that we have to understand naive bayes classifier. It comes under supervised learning algorithms and it is based on Bayes theorem so there is a formula of Bayes theorem.



Bayes theorem is used to determine the probability of a hypothesis with prior knowledge and It depends on the conditional probability. In above formula P(A|B) is Posterior probability **:** Probability of hypothesis A on the observed event B. P(B|A) is Likelihood probability **:** Probability of the evidence given that the probability of a hypothesis is true. P(A) is Prior Probability **:**Probability of hypothesis before observing the evidence and P(B) is Marginal Probability **:** Probability of Evidence.

We took assumption in naive Bayes as the occurrence of a certain feature is independent of the occurrence of other features means each feature individually contributes to identify that it is in which class without depending on each other.

Gaussian naive Bayes is an extension of naive bayes and an assumption often taken is that the values associated with each class are distributed according to a normal distribution or gaussian distribution so after analysing our data we found that our data is normally distributed, so we decided to use gaussian naive bayes as our classifier. Now we will split our data in 80:20 format where 80 is the trained data and 20 is the test data that we applied in our classifier.

**Result and Analysis**

The restaurant was assigned a star of 3.5 after the analysis of the test dataset which was very close to the original value of 3 stars.

Withthe help of the confusion matrix that was generated it was observed that the test set was classified with an accuracy of 73.5%. Our model had a precision of 92.23 %.

The accuracy can be improved if the model is fed more than 1000 data points. If you have a large data set like in millions then it’s strongly recommended to take your model as a Gaussian Bayes classifier because it has good accuracy and precision.

**Conclusion**

The restaurant was assigned the correct number of stars with a little deviation. Based on the result of this paper customer satisfaction analysis might be used by the Naive Bayes classification method to learn sentiments of customers, since customer review is essential in terms of the restaurant business. According to the result of this paper the Gaussian Naive Bayes model yields an accuracy of 73.5% .Further research can be done by increasing the number and variety of the review dataset. To increase the value of accuracy other methods can also be implemented.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***