# Sentiment Analysis for the customer reviews on Zomato

By

Bala Kiran, Kiran Sai, Kailash Chowdary

Data Science Master Program

Business Intelligence

Högskolan Dalarna

Borlänge, Sweden

E-Mail: h19baman, h19kirta, h19kaibo

**Abstract- The aim of this paper is to analyze the reviews given by the customer for the restaurants in Bangalore on a public platform-*Zomato*. Firstly, we have analyzed the patterns between the variables *rate* and *cost for two*. Moreover, we also tried to find the best location for a new restaurant based on the dishes they offer.**

Keywords: Business Intelligence; Machine Learning; Prediction, Analysis, Data Cleaning; Naïve Bayes classifier; Sentiment Analysis;

## 1. Introduction

Online food delivery is one of the most loved concepts in today's world. In our busily, entangled lives, the luxury of receiving our preferred choice of food at our doorsteps has almost been life-changing for the masses. Zomato, the industry stalwart whose name not many would miss recalling, has almost become synonymous with quick and comfortable food [1].

The aim of this paper is to measure the sentimental values of the customer reviews on restaurants and services. So, for the sentiment analysis we have developed a model that will analyze the reviews in and based on the training it has received from the training data set it is going to classify the reviews. Sentiment analysis can be positive-negative or multi-class (3 or more classes), so depending on the nature of the study a choice of such sentiment classification is used. Based on our study approach we would be using multi-class sentiment analysis.

In this study we have used classifiers to classify the reviews which produce different classification results. And we also tried to predict the rating of the restaurant using different variables.

Natural Language Processing(NLP) coves a broad range of techniques that apply computational analytical methods to textual content which provide means of categorizing and quantifying text [2].

Sentiment analysis seeks to quantify the emotional intensity of words and phrases within a text. Some sentiment analysis tools can also factor in the emotional weight of other features of language such as punctuation or emojis [2].

The Vader sentiment tools generate positive, negative, and neutral sentiment scores for a given input. VADER (Valence Aware Dictionary and sentiment Reasoner) is a sentiment intensity tool to NLTK. VADER is ready to go for analysis without any special setup. VADER is unique in that it makes fine-tuned distinctions between varying degrees of positivity and negativity [2].

The best classifier for better prediction of the reviews being sentiments i.e positive, negative, or neutral is

- Naïve Bayes classifier

Naïve Bayes classifier is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naïve Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature [2].

There are different types of model using Naïve Bayes in Python. They are:

- Gaussian Naïve Bayes: It is used in classification and it assumes that feature follow a normal distribution.
- Multinomial Naïve Bayes: It is used for discrete counts
- Bernoulli Naïve Bayes: The binomial model is useful if your feature vectors are binary.

For the study only Multinomial Naïve Bayes classifier is used for classification.

For predicting the rating in the dataset study, we have used decision tree algorithm.

Decision Tree is a decision support tool that uses a tree-like graph or model of decision and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements [4].

## II. DATA SET

In this paper we have analyzing the dataset of Zomato which has established with different type of restaurant at different places in Bengaluru aggregate rating of each restaurant. We have taken the dataset from Kaggle [5]. The attributes of the data set are listed as follows:

url, address, name, online_order book_table, rate, votes, phone, location, rest_type, dish_liked, cuisines, approx._cost(for two people) -- average_cost, reviews_list, menu_item, listed_in(type) – restaurant_type, listed_in(city)--locality.

Out of all the 17 attributes listed above we are using different attributes at different points. The some of main important attributes are:

- Online_order – The column has Yes, or No values weather the its an online order or not.
- Rate – The column has float values which is given by the customer as the rating for the restaurant. It values ranges from 1 – 5.
- Location – Its values states the location of the restaurant.
- Restaurant_type – The column values states the type of the restaurant.

Most of the people in the city is ae mainly dependent on the restaurant food as they do not have time to cook for themselves. With such an overwhelming demand of restaurants it has therefore become important to study the demography of the location. So, we can conclude what kind of food is more popular in a locality and we can make some conclusions and help people to open new restaurant by suggesting the cuisines and restaurant type for the success.

## III. METHODOLOGY

### a. Data Handling

Firstly, we have loaded the data into the data frame using python. Then we performed necessary data filtering for the data.

- We have removed the irrelevant attributes and renamed the important attributes.
- Cleaned the required attributes like average_cost and rate for the better performance.
- Removed the null values from the attributes
- Converted average_cost and rate from string to float.

We have removed all the null values from the data frame so we could have a better prediction and sentiment analysis.

### b. Exploratory Data Analysis

We have analyzed what kind of food the customer would prefer to have and type of restaurant they would prefer. In addition, we have analyzed how they would rate a restaurant. And how the cost would affect the rating of the restaurant. The count of the top cuisines at every location.

### c. Natural Language Processing:

Data cleaning is the standard step used in Natural Language Processing (NLP) such as cleaning the text like removing punctuations, stop words, special characters, and word tokenization.

### d. Training and Testing data sets:

In this experiment of predicting sentiments using multinomial Naïve Bayes we have divided 80% of data as training data and 20% of data as testing data.

In the experiment of predicting the rating using the required features of the data set we would be using decision tree algorithm and we have chosen the same ratios of training and testing data splits as mentioned above.

### e. Techniques Performed

We have performed statistical based techniques for model development. The model is developed using the training and the testing data sets. The model will be trained using sample of the training data and would be trying to predict the values in the test sample data. We use accuracy of the model, to check how well the model is trained.

We have used Multinomial Naïve Bayes classifier under SKLEARN package in python for training and testing the model for the sentiment scores. Generally, this technique is used on data that is multinomially disturbed. It is one of the standard classic algorithms. Which is used in classification problem.

We have used Decision Tree under SKLEARN package in python for training and testing the model for prediction of the rating in the data set.

## IV. DATA EXPLORATION

Let us investigate some conclusions in our data set. Let us look the into the online order service and book table service in the dataset.

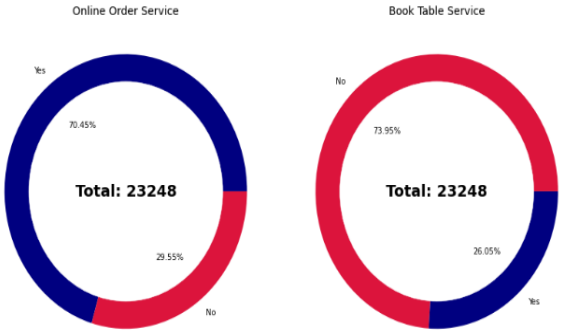


Fig 1: Online order service and Book table service

If we investigate fig 1 70% of the customer would be preferring to order the food online and 26% of customers in the remaining 30% of the customer would like to prefer to book the table in advance.

Let us investigate the percentage of restaurants based on their rating.

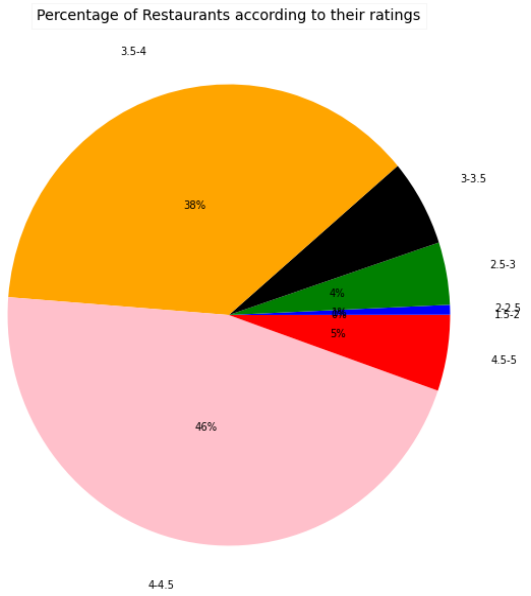


Fig 2: Percentages of Restaurants based on Ratings

From the above plot we can conclude that 46% of the restaurants are been rated between 4-4.5 and only 5% of the restaurants are rated more than 4.5.

Let us investigate into the rest types and check with the count of top 20 restaurant.

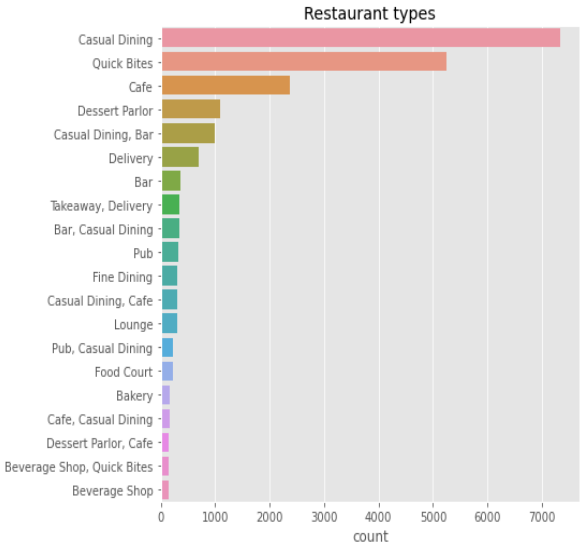


Fig 3: Top 20 Rest Types

The above plot shows the top 20 rest types. And we can conclude that Casual Dining and Quick Bites are the top rest types.

Now let us see the top rating restaurants above 3000 cost.

| | name | votes | rate | average_cost | restaurant_type | location |
|---|---|---|---|---|---|---|
| **18849** | Rim Naam - The Oberoi | 988 | 4.6 | 3000.0 | Buffet | MG Road |
| **19375** | Rim Naam - The Oberoi | 988 | 4.6 | 3000.0 | Dine-out | MG Road |
| **21280** | Rim Naam - The Oberoi | 988 | 4.6 | 3000.0 | Buffet | MG Road |
| **21757** | Rim Naam - The Oberoi | 988 | 4.6 | 3000.0 | Dine-out | MG Road |

Fig 4: Top Rating Restaurants

From the above figure we can see the restaurants which have been rate more than 4.5 and have the average cost of 3000 INR for two people.

Now let us restaurants with top rating and less cost.

| | name | votes | rate | average_cost | restaurant_type | location |
|---|---|---|---|---|---|---|
| **8101** | Milano Ice Cream | 2090 | 4.9 | 400.0 | Desserts | Indiranagar |
| **19004** | Belgian Waffle Factory | 1750 | 4.9 | 400.0 | Delivery | Brigade Road |
| **19325** | Belgian Waffle Factory | 1750 | 4.9 | 400.0 | Desserts | Brigade Road |
| **16957** | Belgian Waffle Factory | 1749 | 4.9 | 400.0 | Delivery | Brigade Road |
| **17239** | Belgian Waffle Factory | 1749 | 4.9 | 400.0 | Desserts | Brigade Road |

Fig 5: Top Rating Restaurants

From the above figure we can see the top restaurants which have been rated more than 4.5 and have low cost less than 500 INR for two people.

## V. RESULTS

Now let us try to predict the rate column using the important features from the dataset. As mentioned earlier, we have trained the model using the training set and tested the model for predicting the values of the test set and checked with the accuracy of the model. We have used Decision Tree as the preferred algorithm. Using the algorithm, we have predicted the rating column with the accuracy of 83%.

Next, we tried to find the sentiment of the reviews given by the customer based on the restaurant type. We have used sentiment analysis tools and raked the sentiment of the reviews and converted the score to positive if it is greater than 0, neutral if it is equal to 0 and negative.

| | name | rating | review | votes | restaurant_type | location | Sentiment Scores | Sentiment |
|---|---|---|---|---|---|---|---|---|
| 0 | Jalsa | 4.0 | beautiful place dine inthe interior take back ... | 775 | Buffet | Banashankari | 0.74 | Positive |
| 1 | Jalsa | 4.0 | dinner family weekday restaurant completely em... | 775 | Buffet | Banashankari | 0.97 | Positive |
| 2 | Jalsa | 2.0 | restaurant near banashankari along office frie... | 775 | Buffet | Banashankari | 0.54 | Positive |
| 3 | Jalsa | 4.0 | went weekend buffet took carte firstly ambienc... | 775 | Buffet | Banashankari | 0.97 | Positive |
| 4 | Jalsa | 5.0 | best thing place ambiance second best thing yu... | 775 | Buffet | Banashankari | 0.95 | Positive |

Fig 6: Sentiment scores

From the above figure 6 we can see that we have predicated the sentiment scores based on the reviews and sentiments based on the sentiment score.

From the above figure, we can see that we have more positive reviews.

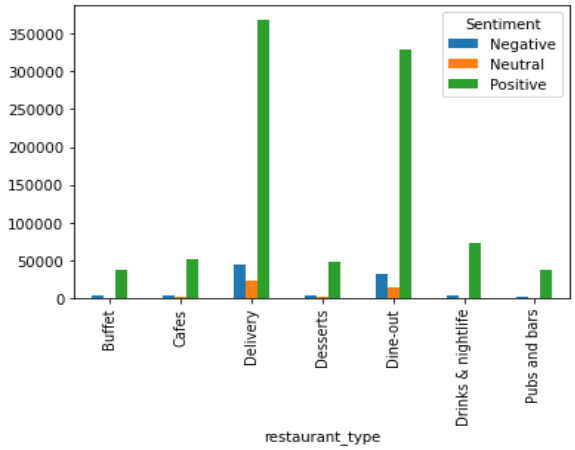


Fig 7: Sentiment Analysis

From the above figure we can conclude that Delivery Restaurant type has the highest reviews and highest positive reviews compare to different restaurant types.

For finding the count of the words we have tokenized the words and removed all the stop words and lemmatized the words.

Doing that we can find the count of the words and we can analyze the highest word which have been used in the reviews.

From the below figure 8 we can see the word counts of the top words which have been used in the reviews. In the reviews people are usually talking more about the place, service, food, ambiance. From those we can assume that having a good place, service and ambiance would be attracting the people towards the restaurant.

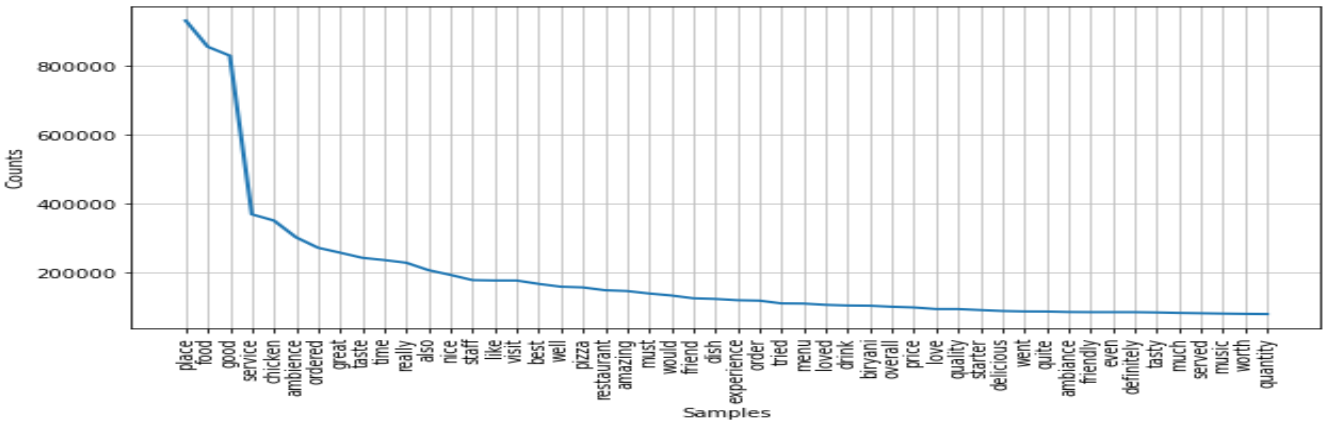


Fig 8: Word Count

Now let us try to predict the sentiment scores using the Multinomial Naïve Bayes and we have achieved an accuracy of 82% for the dataset.

For the better accuracy of prediction, we have transformed the testing data and tried to predict the data and we have achieved the accuracy of 86%.

Based on your study and assumptions, we would like to suggest some locations for new people to open a new restaurant. Like if you would prefer to open a new restaurant in BTM location you can choose North Indian as your top cuisine and choose Paratha as your top dish.

If you are an expect in making Biriyani you can choose Banaswadi, Bannerghatta Road and some preferred locations to open your restaurant. We can also assume that Banaswadi is also a best location to open a new Café or Fast Food cuisine.

And based on restaurant type you could choose your new restaurant to establish at different location based on risk you can choose the common disk as an risk averse and you can go with an new disk to that location as an risk prone preference.

## VI. DISCUSSION

Based on the results and assumptions we tried to find the best restaurant for an new person to open and try to suggest him with an idea so he can act accordingly to his decision of risk prone or risk averse and gave the person with an idea of reviews how does customer behave and how do they respond on the restaurant activities like place, ambiance, service and food. We also tried to predict the rating of the restaurant based on the features available in the dataset and tried to predict the sentiment scores of the reviews given to the restaurant. In both the cases the model has given a reasonable accuracy.

## VII. LIMITATIONS

All the results which have been given would be as assumption based on our analysis and study. It could be different with different perspectives and ideology. We tried to give our assumptions without any biased features. We just tried to study the customers at an particular location and their preferences towards dishes they liked and the rating.

## VIII. REFERENCES


[1] N. Sharma, "Code brew," [Online]. Available: https://www.code-brew.com/blog/2020/02/03/the-unstoppable-zomato-a-success-story-to-watch-out-for-sure/. [Accessed 23 05 2020].

[2] Z. W. Saldaña, "Programming Historian," [Online]. Available: https://programminghistorian.org/en/lessons/sentiment-analysis#exploring-text-with-sentiment-analysis. [Accessed 24 05 2020].

[3] S. Ray, "Analytics Vidhya," [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. [Accessed 24 05 2020].

[4] R. S.Brid, "medium," [Online]. Available: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb. [Accessed 24 05 2020].

[5] H. Poddar, "Kaggle," Kaggle, [Online]. Available: https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants. [Accessed 24 05 2020].