Restaurant Recommendation by Review Analysis on Zomato Dataset

(IBM-312: Data Mining and Business Analytics)

Submitted
by

Shivam Yadav	**20115135**
Sambit Prabhu	**20117108**
Anjani Kumar Tiwari	**20115015**
Rohit Pal	**20115116**

Indian Institute of Technology Roorkee

2023

Abstract

The aim of this project is to develop a model for the prediction of the priority list of restaurants from the available restaurants by taking the dish and budget as customer input. In this project, we analyzed the Zomato Dataset, and using sentiment analysis, text analytics, observing word clouds, DTM, and tf-idf matrices, we have developed a model to recommend the priority list of restaurants for the searched dish and budget. We have used a decision tree classification algorithm, for the classification of restaurants based on available budget, rating of the restaurant, and extra features provided. This model will help the customer to get a priority list of restaurants.

Keywords: Text Analytics, sentiment analysis, DTM, Decision tree classifier.

Introduction

These days due to advancements in technology, businesses are growing at a breakneck pace. The same advancement has been observed in the food industry. The food is something like Its taste matters to everybody's feelings. We usually go to restaurants or order food to celebrate any event or just because of the mood to eat something delicious. And here comes the problem of choosing the best one, in a city there are plenty of restaurants available and obviously we are not aware of all of them. Also, along with the discounts available on the online food ordering system it becomes too hard to decide which one to choose as we want a money-worthed meal. Now here we look at the reviews but again it's not possible to read a maximum of reviews as we want the food at the earliest! So it will be good if we have a system that can provide us with a list of restaurants for a particular meal and budget. In this project, we have analyzed the reviews given by customers for their ordered food.

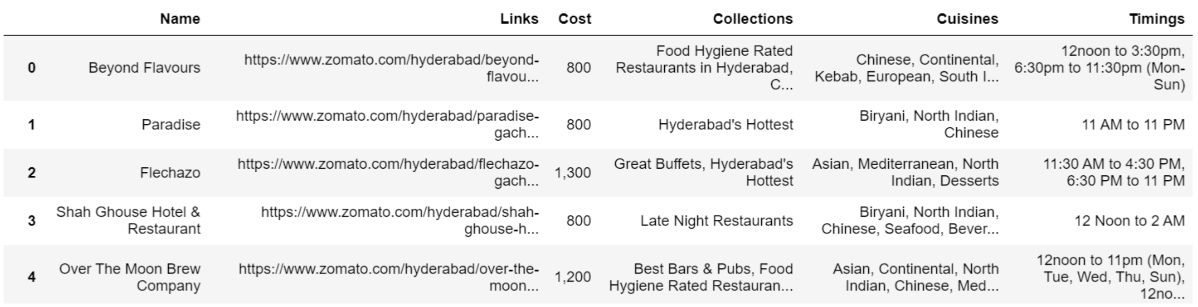
Problem statement

The problem statement contains the review of restaurants connected with Zomato in Hyderabad, the dataset consists of the restaurant's name, available cuisines, and reviews of customers. Now our aim is to make a program that involves text and sentiment analysis of customer reviews and use the obtained values along with the cost of food to provide the user with a priority list of restaurants made with a balance between the cost of food and it's feedback provided so that it becomes easy for a user to pick the best restaurant as per his budget.

Dataset and visualization

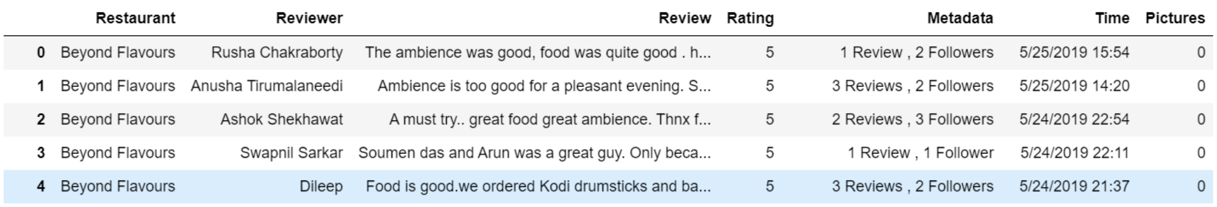
The dataset we have used is Restaurant reviews of Zomato from Kaggle. There are two CSV files, one of them contains the name of restaurants, the average cost of food and cuisines while the other contains 10000 reviews (100 reviews per restaurant), restaurant ratings, and customer names. The data entries are mainly in the form of strings including punctuation signs and emojis. DatasetLink:https://www.kaggle.com/code/ajeetchaudhary/nlp-on-zomato-restaurant-reviews-sentiment-rf-svd/input

Table 1: Dataframe of Restaurant, Cuisines, and Cost Data



	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	800	Hyderabad's Hottest	Biryani, North Indian, Chinese	11 AM to 11 PM
2	Flechazo	https://www.zomato.com/hyderabad/flechazo-gach...	1,300	Great Buffets, Hyderabad's Hottest	Asian, Mediterranean, North Indian, Desserts	11:30 AM to 4:30 PM, 6:30 PM to 11 PM
3	Shah Ghouse Hotel & Restaurant	https://www.zomato.com/hyderabad/shah-ghouse-h...	800	Late Night Restaurants	Biryani, North Indian, Chinese, Seafood, Bever...	12 Noon to 2 AM
4	Over The Moon Brew Company	https://www.zomato.com/hyderabad/over-the-moon...	1,200	Best Bars & Pubs, Food Hygiene Rated Restauran...	Asian, Continental, North Indian, Chinese, Med...	12noon to 11pm (Mon, Tue, Wed, Thu, Sun), 12no...

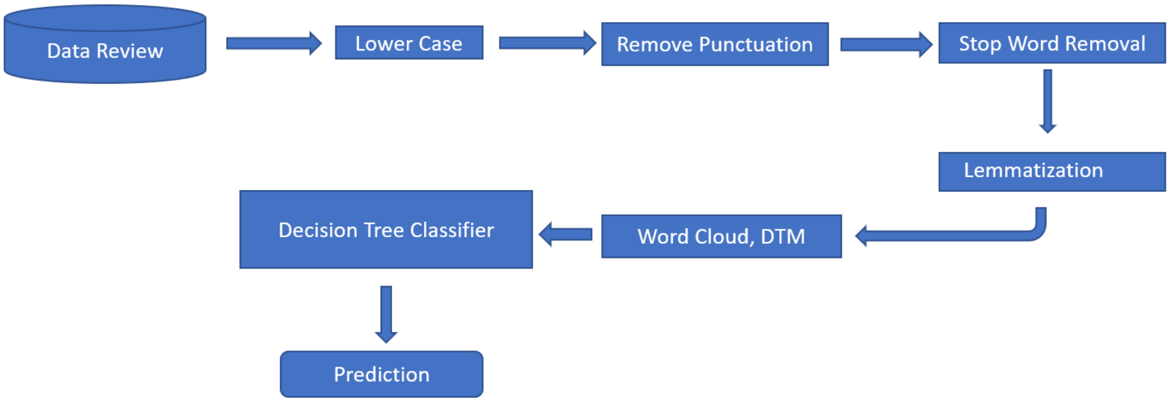
Table 2: Dataframe of Restaurant Review and Ratings Data



	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0

Text Analysis

For drawing conclusions from the review of customers, text analysis is essentially required that will help us to find the most frequent words and phrases used by consumers in their reviews and hence to decide if the service provided was satisfactory or not. For doing the text analysis of the reviews, we have used a for loop that will iterate through the rows of data. In one go we will take 100 rows that correspond to the review of one restaurant (since each restaurant has 100 reviews) and concatenate the entries of all 100 rows together with a space character between them. Now we convert the large combined string to lowercase and tokenize it for further analysis. We have removed punctuations and stopwords available in the nltk library's English dictionary. Now each token has been taken to its original form basically by changing its verb form to the original one (known as lemmatization). So here it completes the part of text cleaning where we have tried to remove some most common and less important words for our analysis. Now we have made the wordcloud of all the available words to check which word is most dominant in number. After that, we have done n-gram analysis of words and observed the frequency of words in reviews of some hotels. By analyzing the bigram and unigram frequency vs words plot we have selected 10 most frequent adjectives and phrases that are most important to decide the quality of service being provided at the restaurant. Based on the frequency of those positive and negative words, the final quality count has been calculated and normalized with respect to all the values to generate the quality index of that restaurant.

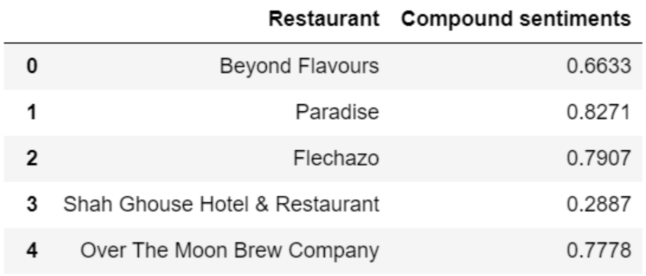


Flowchart for Text Analysis

Sentiment Analysis

We have also calculated the sentiment analysis of the reviews towards a restaurant and its food, which can be used to rate the restaurant and its price of orders. SentimentIntensityAnalyzer from nltk is used for calculating the sentiments. This has four types of scores about the sentiment on any statement neg, neu, pos, and compound. We have taken the Compound Sentiment score.

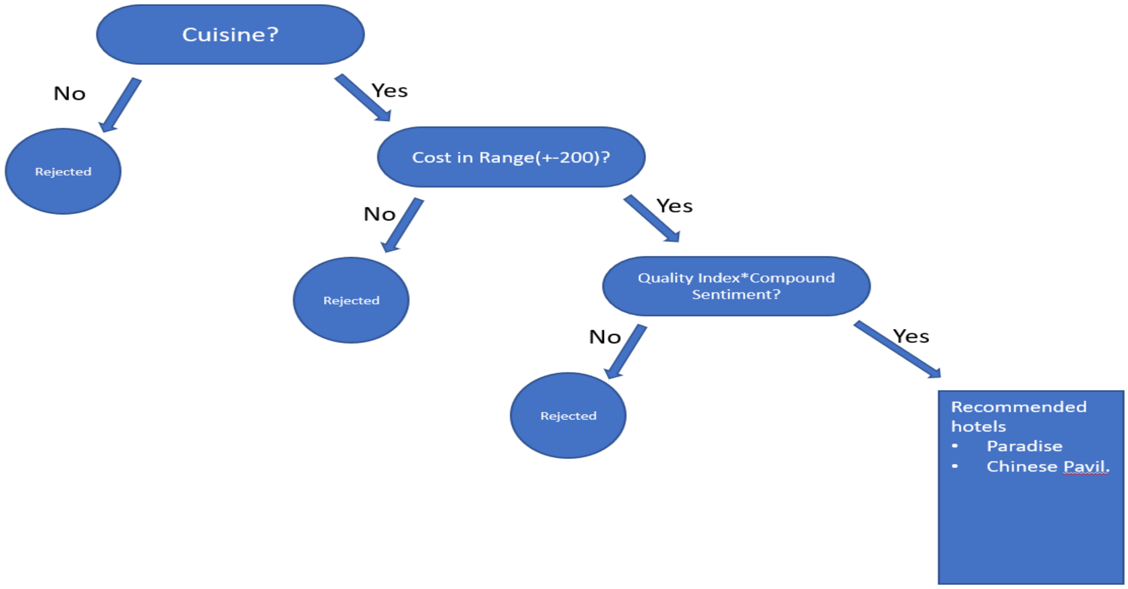
Table 3: Compound Sentiment Analysis Restaurant Review



	Restaurant	Compound sentiments
0	Beyond Flavours	0.6633
1	Paradise	0.8271
2	Flechazo	0.7907
3	Shah Ghouse Hotel & Restaurant	0.2887
4	Over The Moon Brew Company	0.7778

Recommendation Prediction

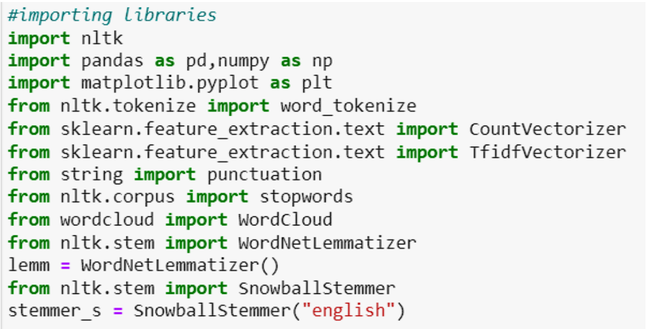
After getting the numerical estimates of compound sentiment and text analysis of reviews it's time to use them for the main task which is predicting the priority list of restaurants given the name of the dish and the budget amount from the user's end. So here we read the cuisine entered by the user and check if it is available in the dictionary or not. Restaurants that have the desired cuisine are promoted to the next level rest are rejected. Now we check if the average cost of food comes within the range of (user's budget +/- 200). Restaurants that satisfy the condition are checked for the next level and the rest of them are rejected. Finally, we arrange the remaining restaurants based on the product of their quality index value obtained from text analysis and compound sentiment value obtained from sentiment analysis. Since a good review will have a high-quality index and if it's a positive review then it will have a significant positive value of compound sentiment, we can use products of both values to compare restaurants in terms of feedback. In the final list restaurants having higher value of product of quality index and compound sentiment will be on the upper side.



Recommendation Prediction Algorithm

Libraries & Modules Used

- **Panda**
- **NumPy**
- **Matplotlib**
- **nltk**
- **sklearn**
- **WordCloud**
- **operator**



```
#importing libraries
import nltk
import pandas as pd,numpy as np
import matplotlib.pyplot as plt
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from string import punctuation
from nltk.corpus import stopwords
from wordcloud import WordCloud
from nltk.stem import WordNetLemmatizer
lemm = WordNetLemmatizer()
from nltk.stem import SnowballStemmer
stemmer_s = SnowballStemmer("english")
```

Results and Analysis

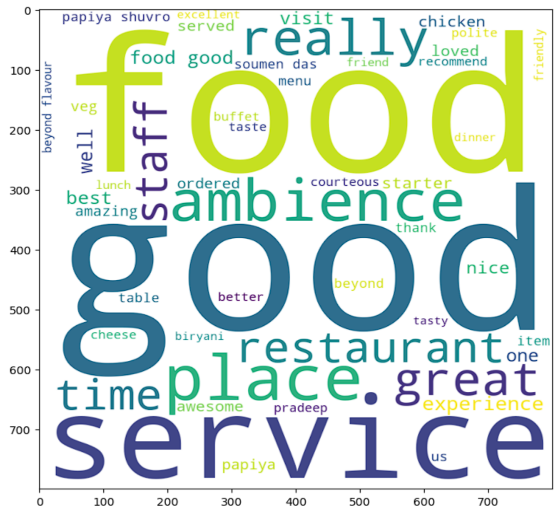


Fig 1: Word Cloud of Beyond Flavours's Reviews

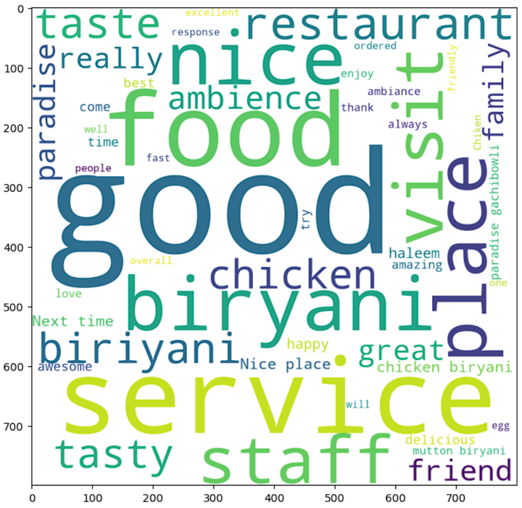


Fig 2: Word Cloud of Paradise's Reviews

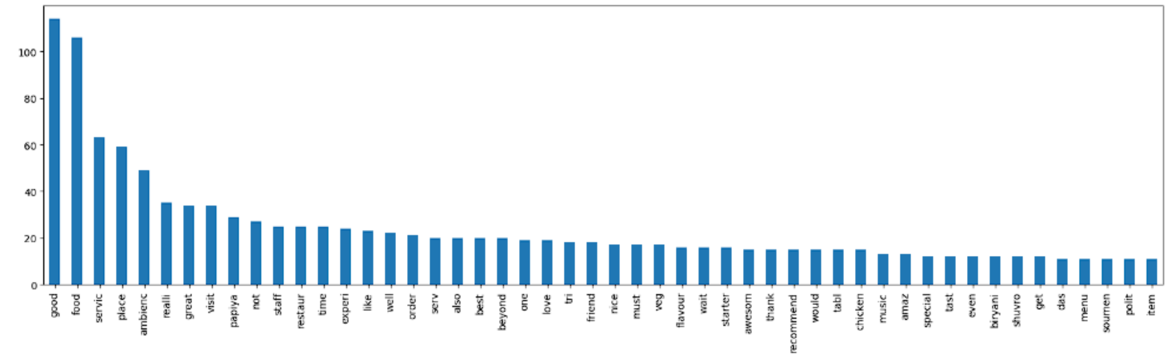


Fig 3: Unigram Analysis of Beyond Flavours's Reviews

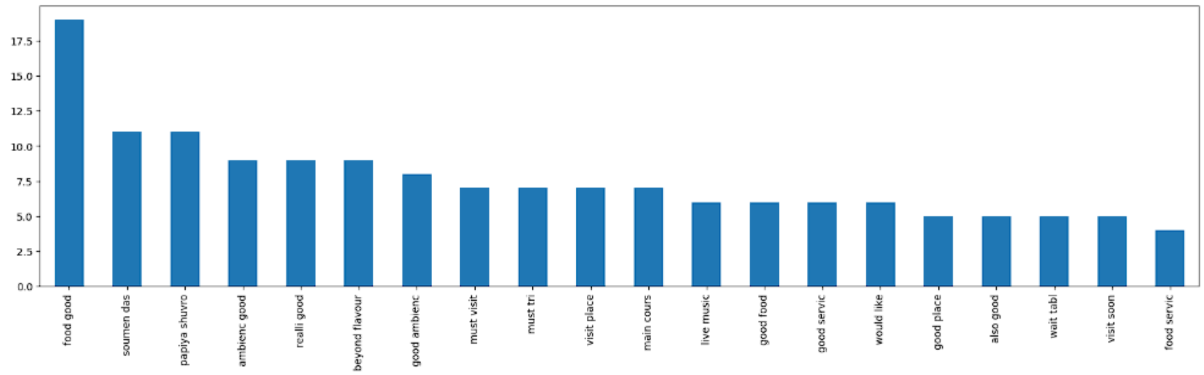
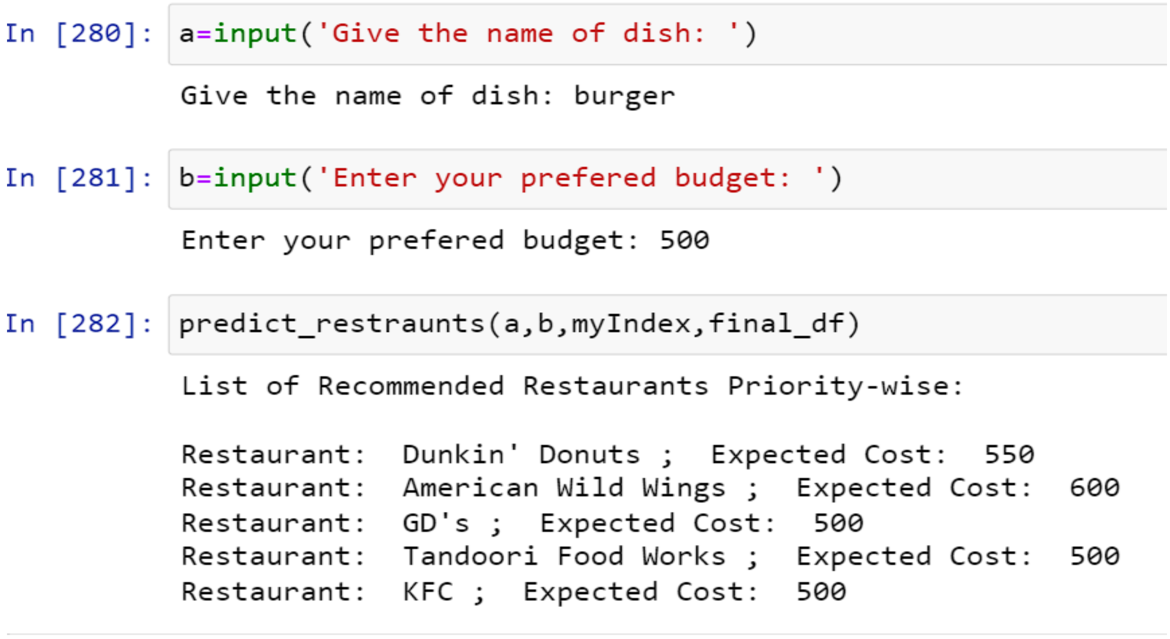


Fig 4: Bigram Analysis of Beyond Flavours's Reviews



```
In [280]: a=input('Give the name of dish: ')

          Give the name of dish: burger

In [281]: b=input('Enter your prefered budget: ')

          Enter your prefered budget: 500

In [282]: predict_restraunts(a,b,myIndex,final_df)

          List of Recommended Restaurants Priority-wise:

          Restaurant:  Dunkin' Donuts ;  Expected Cost:  550
          Restaurant:  American Wild Wings ;  Expected Cost:  600
          Restaurant:  GD's ;  Expected Cost:  500
          Restaurant:  Tandoori Food Works ;  Expected Cost:  500
          Restaurant:  KFC ;  Expected Cost:  500
```

Fig 5: Final Output List of Restaurants Recommended

The above figures show the results of various pieces of code to analyze the text. Fig.1 & 2 show the wordcloud of reviews of two different restaurants while Fig. 3 & 4 show the bar graph of words vs frequency for n-gram analysis (bigram & unigram) by taking one word at a time and taking two words at a time. It can be clearly seen that word “good” is dominant in the reviews of “Beyond Flavours” restaurant also the bigram has a very high frequency of positive words that depicts that the overall response is good for this restaurant.

Fig. 5 shows the recommended result for the dish “burger” and a preferred budget of 500 rs. In the predicted priority list Dunkin’ Donuts is predicted as the best restaurant to visit in slightly above price but better quality.

Interesting findings/insights/observations

- In this project, we found that customers are very happy with the quality of food of some restaurants even though the cost is higher.
- Peace, hygiene, and staff interaction also matter to them.
- The average price of the order is roughly Rs 861.42.
- Chicken Biryani is the favorite and most reviewed product.

Conclusion

In this project, we have developed a classification model, which suggests restaurants from which the customer should make an order on the basis of his/her budget, product availability, and quality of food suggested by the other customers. In this project, we have used sentiment analysis, text analytics, observing word clouds, DTM, and tf-idf matrices. We have also used the decision tree classification algorithm. It will be very useful for food lovers to get good suggestions without any type of misleading by some other person. In the future, this model will be much more helpful for analyzing the data and overall development of businesses. In the future, we can consider the location of restaurants, traffic, and traveling time to the restaurant which will definitely improve our model performance.

References

1. Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants Using Random Forest Classifier https://shorturl.at/dABD6

2.https://www.nltk.org/

3. Alhadethy, Mohanad & Hamad, Mortadha & Adnan Jaleel, Refed. (2021). Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes. Systems Analysis Modelling Simulation.