**Mazin Kamaleldin Hassan Abdalla**

**Student #: 501115000**

**Supervisor: Ceni Babaoglu, PhD**

**07/03/2022**

This research uses machine learning and natural language processing (sentiment and emotions analysis) to answer the following questions:

1. What is the overall customer sentiment about the restaurant?
2. What types of emotions are customers displaying while they are writing their reviews?
3. And apart from the served food quality, what other restaurant attributes affect the overall customer rating?

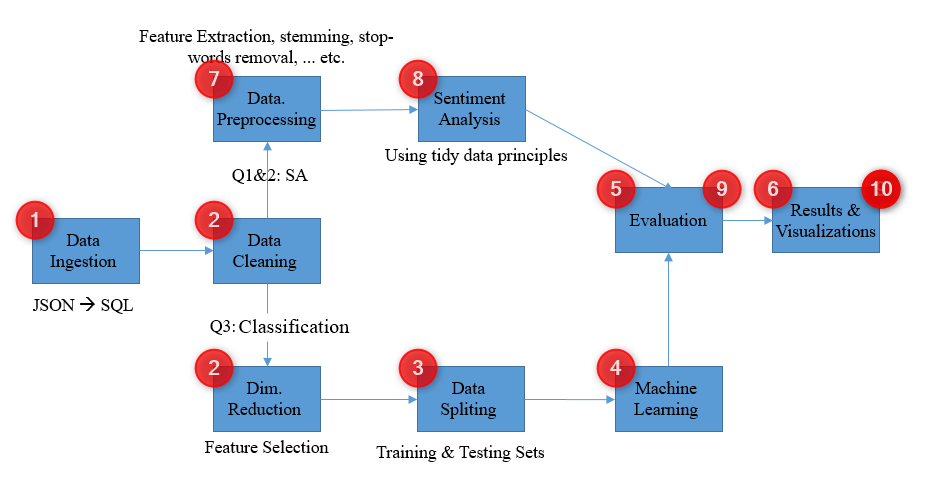


Figure : Tentative Methodology

In this file, we will follow the numbers on the tentative methodology above to explain the R code associated with this phase of the project (Initial Results & Code.RMD). For this submission, the [GitHub](https://github.com/mazinkamal134/CIND820) repository has the following files:

|  |  |
| --- | --- |
| **File Name** | **Description** |
| [Businesses.csv](https://github.com/mazinkamal134/CIND820/blob/main/Businesses.csv) | A text file containing restaurants’ data in a tabular format. |
| [Reviews.csv](https://github.com/mazinkamal134/CIND820/blob/main/Reviews.csv) | A text file containing reviews data in a tabular format (a sample of 10,000 reviews) |
| [Descriptive Stats.RMD](https://github.com/mazinkamal134/CIND820/blob/main/Descriptive%20Stats.RMD) | R script file, containing the descriptive statistics conducted during the literature review of this project. |
| [Initial Results and Code.Rmd](https://github.com/mazinkamal134/CIND820/blob/main/Initial%20Results%20and%20Code.Rmd) | R markdown file containing the code for data analysis, data cleaning, machine learning, and natural language processing for both sentiment and emotion analysis. |

Table : GitHub Repository Files

**Steps**:

Question 3: Classification

1. As discussed in the previous chapter, Yelp data (businesses and reviews) comes in JSON format, where SQL Server OPEN\_JSON function was utilized to ingest the JSON data into a tabular format. In our R code, the data is read directly from SQL Server into a data frame, where all further analysis is done using R. An alternative ingestion method is also provided, where the data frames are created by reading the CSV files directly from the GitHub repository.
2. Review data is clean, no cleaning or dimensionality reduction processes will be applied to this data set, however, it will directly be analyzed using tidytext and dyplr functions. On the other hand, business data needs cleaning and feature reduction. The following steps were used to put the data in a final shape suitable for machine learning:
   1. Removing attributes that are not relevant to this analysis, such as business name and address details (no geo-analysis is being done here), in addition to attributes we used to filter out the restaurants’ data, such as “Is Open” and “Categories”.
   2. Removing data columns with too many missing values where any attribute with more than 25% of missing values will be ignored, attributes such as: “Wheelchair Accessible”, “Dogs Allowed”, “Accepts Bitcoin”, ... etc. Proper comments are provided in the RMD file describing why a certain attribute was either ignored or included in the analysis.
   3. Reducing highly correlated columns, fortunately, the few remaining numeric columns showed almost no correlation between them, so they were all kept as part of the dataset. All other attributes are logical describing the observations by True or False.
3. Before splitting the data into training and testing data sets, a problem had to be addressed. The data across the response variable is severely imbalanced, where a significant amount of star ratings is reported for 3, 4, and 5 stars, with very few ones reported against 1-, and 2-star ratings. So over-sampling was used to balance the data set around the class label (star rating). Then the over-sampled data set was split into training and testing data sets.
4. Two algorithms were used to classify the restaurants’ data (new over-sampled data set) to answer question 3 of this research; Support Vector Machine and Random Forest classifiers. 10-folds cross-validation was used to generate two vectors of each algorithm’s accuracy across the different folds.
5. The accuracy vectors from #4 above were analyzed using ANOVA, where it turned out that SVM performed differently from Random Forest, and since the RF displayed a higher accuracy across all data folds, it was chosen to deduce the most important attributes contributing to tree(s) construction, i.e., attributes that are affecting the restaurant's overall star rating. RF was trained again using the full over-sampled training data set yielding an accuracy of 76%, then the importance function of “randomForest” package was used to calculate the Mean Decrease Accuracy & Mean Decrease Gini as to be able to identify the most important attributes, also “varImpPlot” function was used to graphically display the variables and their importance, as below:

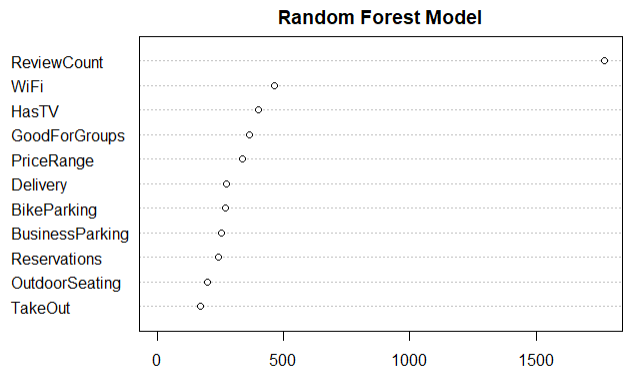


Figure : Random Forest Variable Importance

Where the number of reviews is the most important variable in determining the restaurants' overall star rating.

1. Various visualizations were provided to interpret and better understand the data while it is being analyzed.

Questions 1 & 2: Sentiment & Emotions Analysis

1. “tidytext” and “dyplr” packages were used to tokenize and clean the review text, using functions such as unnest\_tokens and anti\_join to remove the stop words. anti\_join was used again to remove the numeric unigrams from the review text, leaving us with clean lower-case tokens ready to be analyzed using the available lexicons.
2. “tidytext” package provides 3 different unigram lexicons, Bing et al, NRC, and AFINN, each with its own set of english words and fatures. For instance, AFINN provides a sentiment score ranging from -5 to +5, while Bing et al tagges each unigram with “positive” or “negative”. Like Bing, NRC lexicon provides positive/negative sentiments, in addition to 8 different emotions associated with each unigram, such as anger, fear, and joy. The 3 lexicons were used to extract the sentiment from the review text, while the NRC alone was used to analyse the emotions associated with reviews corresponding to each star-rating group.
3. A correlation test comparing the sentiment score obtained from each lexicon against the review star rating was carried out. Both Bing et al and AFINN showed some correlation with the actual data, while NRC did the opposite, unfortunately. Bing et al lexicon was then chosen to provide further analysis and draw conclusions about the data. Using Bing alone, we were able to show that the sentiment across 1 and 2-star ratings tended to be negative, while for a 3-star rating both positive and negative sentiments were seen, however, 4 and 5-star ratings showed clearly positive sentiments. We also found that the average sentiment score was negative for both 1 and 2-star rating reviews, while it was positive for 3, 4, and 5-star ratings with the latter scoring the highest score among all. On the other hand, when emotions were analyzed, 1 and 2-star rating reviews displayed emotions such as anger and disgust, while the other ratings displayed emotions such as trust and joy.
4. Various visualizations were provided to interpret and better understand the data while it is being analyzed. The most interesting visualization in this section was the word cloud, which was generated for each rating group directly from the tokenized data (tidytext). Words such as “food” and “service” were dominant across all rating groups with different contexts of course, while words such as “worst”, “bad”, and “disappointed” were evident across 1 and 2-star rated reviews. In contrast, words such as "nice”, “amazing”, and “awesome” were seen across reviews with 3, 4, and 5 stars.