

Restaurants: A Comprehensive Analysis

By

**Manmeet Kaur Sahota**

A dissertation submitted for the degree of

M. Sc. Data Analytics

Dublin Business School

August 2020

Supervisor: **Mr. Richard O’Callaghan**

# Declaration:

I, Manmeet Kaur Sahota, declare that this research is completed myself for the award of Masters of Science in Data Analytics.

Moreover, the work on this project is complete self-work and analyses and I have not copied from any other site except for where the previous work is cited and acknowledged properly within the text. This work is completely amenable with the Dublin Business School’s guidelines.

Manmeet Kaur Sahota

Date: 25th August 2020

I am the shoe he we then for me at this very place for another hour may be then she will realise that she will not be right however

# **Acknowledgement**

I would like to thank my supervisor, Richard O’Callaghan, for his tremendous support and understanding. He guided me through my research work and provided immense technical knowledge and feedback which greatly enhanced my project.

I would also like to thank my family and friends for encouraging and motivating me throughout the journey.

I am the he she me they were of this there at me

# Abstract:

Online customer ratings are valuable tools for businesses to enhance their service quality and for customers to collect information for decision-making prior to their service purchase. Determining the customer impression is considered one of the prominent factors on the success of the restaurant businesses. Due to the rapid growth of digital contents related to restaurant or foods in the web, people are more inclined on reviews before going to any restaurant so the significance of customer review is inevitable. In real-life scenarios, these reviews can be very useful for restaurants to maintain and improve their standards based on the findings. In this project, we are analysing customer reviews available through Yelp dataset. It consists of customer ratings and textual reviews which are being used in sentimental analyses and predicting the star ratings of restaurants based on supervised machine learning algorithm such as Support Vector Machine. We are also calculating polarity of positive and negative words to gauge the best and to be improved attributes of the restaurants. We are examining user reviews on the restaurant environment, service and price and food quality. We have found that users have found the service and food quality as the good services which need improvisation for some restaurants whereas environment stands at a lower priority.

# List of Figures:

[Figure 1: Steps to analyse data 16](#_Toc49710839)

[Figure 2: Project Plan 17](#_Toc49710840)

[Figure 3: Block Diagram of Web Scraping 18](#_Toc49710841)

[Figure 4: Web Scraping Output 19](#_Toc49710842)

[Figure 5: List of Files in Yelp Dataset 19](#_Toc49710843)

[Figure 6: Description of Business JSON 20](#_Toc49710844)

[Figure 7: Business Dataset 20](#_Toc49710845)

[Figure 8: Description of Review JSON 21](#_Toc49710846)

[Figure 9: Review Dataset 21](#_Toc49710847)

[Figure 10: Description of Photo JSON 21](#_Toc49710848)

[Figure 11: Description of the Tip JSON 21](#_Toc49710849)

[Figure 12: Description of the User JSON 22](#_Toc49710850)

[Figure 13: Raw Business Data 23](#_Toc49710851)

[Figure 14: Data Filtering 23](#_Toc49710852)

[Figure 15: Restaurant Selection 23](#_Toc49710853)

[Figure 16: Cleaned Business Dataset 24](#_Toc49710854)

[Figure 17: Unclean Review Dataset 24](#_Toc49710855)

[Figure 18: Cleaned Review Dataset 24](#_Toc49710856)

[Figure 19: Merged Data Part 1 25](#_Toc49710857)

[Figure 20: Merged Data Part 2 25](#_Toc49710858)

[Figure 21: Creating Labels column and removing special characters from review text 26](#_Toc49710859)

[Figure 22: Restaurants by Category 27](#_Toc49710860)

[Figure 23: Restaurant by State 28](#_Toc49710861)

[Figure 24: Restaurants by City 28](#_Toc49710862)

[Figure 25: Reviews by Cuisine 29](#_Toc49710863)

[Figure 26: Top 5 States with most reviews 29](#_Toc49710864)

[Figure 27: Top 10 cities with Most Reviews 30](#_Toc49710865)

[Figure 28: Top 10 Restaurants with most reviews 30](#_Toc49710866)

[Figure 29: Top 10 restaurants with American Cuisine 31](#_Toc49710867)

[Figure 30: Count of Restaurants against Ratings 31](#_Toc49710868)

[Figure 31: Percentage of Positive Reviews by Cuisine type 32](#_Toc49710869)

[Figure 32: Average rating of each category 32](#_Toc49710870)

[Figure 33: Heat map of positive and negative words in Review 33](#_Toc49710871)

[Figure 34: Distribution of Review Stars 33](#_Toc49710872)

[Figure 35: Model Data in SVM (Bishop, 2006) 34](#_Toc49710873)

[Figure 36: Output of SVC Model 35](#_Toc49710874)

[Figure 37: Words in American Restaurant 36](#_Toc49710875)

[Figure 38: Actual VS Predicted Ratings 37](#_Toc49710876)

[Figure 39: Graph of Predicted VS Actual Ratings 37](#_Toc49710877)

[Figure 40: Classification Report 37](#_Toc49710878)

[Figure 41: Confusion Matrix 38](#_Toc49710879)

[Figure 42: Top 15 positive words for The Arrogant Butcher 39](#_Toc49710880)

[Figure 43: Top 15 negative words for The Arrogant Butcher 39](#_Toc49710881)

[Figure 44: Graph for Top Words for Restaurant 1 40](#_Toc49710882)

[Figure 45: Top Positive Words for Maestro’s Ocean Club 41](#_Toc49710883)

[Figure 46: Top Negative Words for the Maestro's Ocean Club 41](#_Toc49710884)

[Figure 47: Graph of Top words for Restaurant 2 41](#_Toc49710885)

[Figure 48: Top Positive Words for StripSteak 42](#_Toc49710886)

[Figure 49: Top Negative Words for StripSteak 42](#_Toc49710887)

[Figure 50: Top Words for Restaurant 3 43](#_Toc49710888)

[Figure 51: Top Positive Words for Windsor 43](#_Toc49710889)

[Figure 52: Top Negative Words for Windsor 43](#_Toc49710890)

[Figure 53: Top Words for Restaurant 4 44](#_Toc49710891)

[Figure 54: Top positive words for Lawry's 45](#_Toc49710892)

[Figure 55: Top Negative Words for Lawry's 45](#_Toc49710893)

[Figure 56: Graph of Top Words for Restaurant 45](#_Toc49710894)

[Figure 57: Top Positive Words for Blue Martini Lounge 46](#_Toc49710895)

[Figure 58: Top Negative Words for Blue Martini Lounge 46](#_Toc49710896)

[Figure 59: Graph of top words for Restaurant 6 46](#_Toc49710897)

[Figure 60: Top Positive words for Bahama Breeze 47](#_Toc49710898)

[Figure 61: Top Negative words for Bahama Breeze 47](#_Toc49710899)

[Figure 62: Top Words for Restaurant 7 48](#_Toc49710900)

[Figure 63: Top Positive Words for Beckett's Table 49](#_Toc49710901)

[Figure 64: Top Negative Words for Beckett's Table 49](#_Toc49710902)

[Figure 65: Graph for Top Words for Restaurant 49](#_Toc49710903)

[Figure 66: Top Positive Words for Egg Works 50](#_Toc49710904)

[Figure 67: Top Negative Words for Egg Works 50](#_Toc49710905)

[Figure 68: Graph of Top words for Restaurant 9 50](#_Toc49710906)

[Figure 69: Top Positive Words for Worth Takeaway 51](#_Toc49710907)

[Figure 70: Top Negative Words for Worth Takeaway 51](#_Toc49710908)

[Figure 71: Graph of top words for restaurant 10 52](#_Toc49710909)

[Figure 72: Matrix of Top Positive Words for American Restaurants 52](#_Toc49710910)

[Figure 73: Matrix of Negative Words for American Restaurants 53](#_Toc49710911)

[Figure 74: Top 10 Negative Words in American Cuisine 53](#_Toc49710912)

[Figure 75: Top 10 Positive words for American Cuisine 53](#_Toc49710913)

Table of Contents

[Declaration: 2](#_Toc49710914)

[**Acknowledgement** 3](#_Toc49710915)

[Abstract: 4](#_Toc49710916)

[List of Figures: 5](#_Toc49710917)

[1. Introduction 9](#_Toc49710918)

[2. Literature Review: 11](#_Toc49710919)

[3. Objectives: 15](#_Toc49710920)

[3.1 Business Problem: 15](#_Toc49710921)

[3.2 Research Question: 15](#_Toc49710922)

[3.3 Aim: 15](#_Toc49710923)

[3.4 Scope: 15](#_Toc49710924)

[3.5 Limitations: 15](#_Toc49710925)

[3.6 Roadmap 15](#_Toc49710926)

[4. Methodology: 17](#_Toc49710927)

[4.1 Data Acquisition: 18](#_Toc49710928)

[4.2 About the Dataset: 19](#_Toc49710929)

[4.3 Data Cleaning: 22](#_Toc49710930)

[5. Exploratory Data Analysis (EDA): 27](#_Toc49710931)

[6. Modelling the Data: 34](#_Toc49710932)

[7. Results and Findings: 39](#_Toc49710933)

[1. The Arrogant Butcher: 39](#_Toc49710934)

[2. Mastro's Ocean Club 40](#_Toc49710935)

[3. Strip steak 42](#_Toc49710936)

[4. Windsor 43](#_Toc49710937)

[5. Lawry’s The prime rib 44](#_Toc49710938)

[6. Blue martini lounge 46](#_Toc49710939)

[7. Bahama Breeze 47](#_Toc49710940)

[8. Beckett’s Table 48](#_Toc49710941)

[9. Egg Works 50](#_Toc49710942)

[10. Worth Takeaway 51](#_Toc49710943)

[11. Combined Analysis of American Restaurants: 52](#_Toc49710944)

[8. Future Scope 55](#_Toc49710945)

[Bibliography 56](#_Toc49710946)

## Introduction

In this innovation-loving environment, the use of web-based applications has increased exponentially over the last decades, which in turn has led to an efficient increase in business and environmental sales (Hao Tian, 2018). Today, we have a number of websites information on the web that provide us with a diverse range of insightful data. We would just like to thank the smartphone like it’s a human being, before making a trip or even a meal at a destination, first we choose to search the destination and then find out more about it and ultimately decide to choose whether or not schedule a visit to this destination. This decision is based on the reviews that he goes through which are posted by various customers who have already taken the services provided by this restaurants and convey their experiences (Jamie, 2012). Due to numerous customer reviews, it becomes difficult for restaurants to examine the customers experience in an up-to date and complete manner.

The 21st century has seen the largest impact of social media on consumer behaviour, which has resulted in massive variations in the way the buyer acts in alignment with each product, and have ranged based on the feedback provided by multiple other users that have already bought the product. (Jeong, n.d.). The social media has been the most powerful medium of connection for companies around the world to communicate with customers. With social connections, within general demographics, media are rising rapidly, many companies have noticed the potential of social media, and their marketing tactics have changed to take advantage of these new opportunities. But social media makes it possible for users to share their purchasing experiences through electronic word of mouth (eWOM) to create a reliable source for other consumers (Yang, 2017).

Online reviews today have the power to connect the potential consumer directly with a restaurant even before he/she walks through the doors of a restaurant (Yang, 2017). Moreover, the popularity of online review websites’ (e.g., Yelp.com, TripAdvisor, and Angie’s List) have increased in recent years, and thus more reviews have been created for a variety of products and services. For instance, Yelp.com has helped consumers select local organizations, such as restaurant, bars, etc. with displaying shared information on dining experiences, (Luca, 2016) and it presents an opportunity to its users to reach the information quickly and easily regarding many restaurants. Furthermore, Yelp.com had 142 million visitors in 2015 and 2.1 million various organizations registered on the website. It is also noted by Luca that this indicated rather strongly that these online review websites could potentially be rich sources of information and thus have helped to decrease the percentage of poor purchasing experiences for consumers.

Machine Learning is becoming an important sector where it is beneficial to find trends and discover unseen patterns in a large array of data. Through applying careful research, we can better define industry patterns, customer interaction habits, marketing trends, advertising performance, and customer satisfaction, and adapt our company to those needs through concentrating on the field that lacks market (Ian Goodfellow, 2016).

Sentiment analysis (SA), a process which combines Natural Language Processing (NLP) and Machine Learning algorithms, which combines the weight of various sentiments across a sentence or document for various entities, factors, product. The sentiment can be categorized a positive, negative or neutral. (Lexalytics, 2020). There are various processes to extract, refine and look for objectives in text which include opinions, reviews and sentiments (Keith Norambuena, 2019).

## Literature Review:

Bo Pang and Lee were the pioneers in this field (Sunil B. Mane, n.d.). Current works include mathematical expressions to evaluate sentences based on the proximity to adjectives and adverbs. Various mechanism has been implemented until now, which includes bags of words, training corpus, document level, sentence level, and feature-level opinion mining . Different polarity measures exist according to the external system wherein sentimental analysis is utilized. (Sunil B. Mane, n.d.)

A range of different feature representations, discussed in details below, have been tested for multimodal sentiment detection; for text detection, a variety of approaches have been explored including lexicon polarity, text2vec and bag-of-words representations. (Amasyali, et al., 2018).

In the 18th century, Reverend Thomas Bayes developed a method known as Naive Bayes that used probability and opportunity approaches (Rachmawan Adi Laksono, 2019). Naive Bayes calculates future probability predictions from data or experiences that have been given, based on the opportunity point of view (Van Dynk, 2007). One characteristic of the Naive Bayes Classification is the existence of independent input variables which assume the presence of an articular feature from a class that is mutually independent of other features (Rachmawan Adi Laksono, 2019). Hence, methods of precision, recall, and accuracy is used to check the accuracy of the results of the process. A confusion matrix is created to provide performance classification data where elements of confusion matrix are True positive (TP) when both human and method predict are positive and True Negative (TN) when both human and method predict are negative (Rachmawan Adi Laksono, 2019). False negative (FN) is used when the human prediction is positive while method prediction is negative and False Positive (FP) is used when the human prediction is negative while method prediction is positive (Rachmawan Adi Laksono, 2019). The level of accuracy between what the user wants and the results of the system process is called Precision whereas Recall is the average success of the system in the process of finding information (Rachmawan Adi Laksono, 2019). Precision and Recall calculations are used to avoid measurement errors for deviation values (Rachmawan Adi Laksono, 2019). Accuracy is the degree of truth between the predictive value and the actual value. Precision value is obtained by dividing TP with the positive results obtained by the method (Rachmawan Adi Laksono, 2019). In the above research, customer satisfaction was calculated using Naïve Bayes Classifier which had limitations of accuracy and prediction of false positive. It had a lot of pre-processing steps to follow in order to get meanings from words (Rachmawan Adi Laksono, 2019).

Incremental training can be used on stream data where training data arrive after a certain timeline (Doan, 2016). In this paper, Jugal constructed an incremental decision tree but the downside of it is that it may result in an unfeasible split resulting in an unstable tree (Doan, 2016). As explained earlier in 9, Online Naïve Bayes classifier uses updated counters to compute the probability of the class to which new training entry belongs to. It computes the posterior probability and use them to adjust the degree of error between classification prediction and observation (Doan, 2016). Whereas for some research, Factorization Machine, a variant stochastic gradient descent has also been used that gives better performance and works well with sparse data where it utilizes regularization in the training set. (Doan, 2016). When Factorization Machine is dealt with polynomial kernel it is often comparable to Support Vector Machines (Doan, 2016).

In a paper written by Min Zhao, using sentiment-oriented pointwise mutual information (SO-PMI) to judge the sentimental word’s polarity and its intensity (Min Zhao & Chai, 2015). SO-PMI is an unsupervised method, which can measure the sematic relevance of words by using the information of words’ co-occurrence (Min Zhao & Chai, 2015). If the probability of two words’ co-occurrence in a certain range of the text is big, then they have a strong correlation. Conversely, correlation of the two words is weak if the co-occurrence is sparse. (Min Zhao & Chai, 2015). In this method, the dependency is basically dependant on the co-occurrence of the words in sentences which can be hugely improvised using the lexicon-based technique (Min Zhao & Chai, 2015).

A sentiment Lexicon is made of index of sentiment words which helps in improving the accuracy by storing polarity information of the relevant word irrespective of the polarity that it carries. (Hanhoon Kang, 2012) . In this process, a lexicon is constructed using sentiment words in restaurant reviews and polarity of the word is found. For further intense analyses, bi-gram is constructed with the help of the adverb (Hanhoon Kang, 2012). Small lexicons were constructed to be used in movie reviews and newspapers to classify words into good or bad (Y, 2008). This became a generic method sooner and hybrid approaches were made in order to classify and rate the words polarity for in-depth analyses (Y, 2008).

The lexicon-based technique computes the sentiment by comparing the input data with a dictionary consisting of pre classified terms (Athindran, et al., 2020). The machine learning approach employs one of the many machine learning algorithms to compute the sentiment (Athindran, et al., 2020). As discussed in this paper by Athindran, the hybrid approach is based on the combination of the machine learning approach and the lexicon based sentiment computation approach, which will have an edge over the individual techniques. Further discussed in this research, the data taken from Twitter for the two new smartphone releases which was crunched, filtered and then fed as the input data to a hybrid model sentiment classifier, where the combination of the Lexicon based approach and the popular Naive Bayes algorithm was employed, which enhanced the accuracy of the model and also ensured robustness of the classification (Athindran, et al., 2020). For comparison, the tweets for each phone were taken along with their features such as camera and display and processed through the hybrid model (Athindran, et al., 2020). This led to establishment of word Cloud for the manufacturer to identify the commonly discussed topics relating to their product by the customers (Athindran, et al., 2020).

Sunidhi Dwivedi proposed an unsupervised dictionary based approach, which discusses about how document level opinion mining system performs that is, data pre-processing - remove unwanted data, sentence split, tokenization, POS tagging and stemming, Extraction of opinion words - adjectives are extracted using Apache Open NLP tool and extracted adjective stored in text file, Seed list preparation- extracted adjectives are matched with the seed list words to find their polarity, Polarity detection, calculation of sentimental score. It compares the sentimental score of different mobiles (Sonawane & Kulkarni, 2017).

Language specific sentimental analysis has also been achieved using the Lexicon based approach and machine learning based approach by (Aye, 2017). In this paper, Lexicon based techniques, a sentiment dictionary with sentiment words is used for sentiment classification, were used efficiently. The dictionary contains polarity of each word whether they are positive, negative and objective words (Aye, 2017). Polarity of the opinion words can be determined by matching those words with dictionary words (Aye, 2017).

Furthermore, an aspect-based opinion polling system takes a collection of text reviews and some predefined aspects as its data, and determines the polarity of each aspect from each review to generate an opinion poll. (Jingbo Zhu, 2011). Jingbo proposes an automatic method of aspect-based opinion polling from unlabelled textual customer reviews. A multi-aspect bootstrapping method is proposed to learn aspect-related terms of each aspect to be used for aspect identification (Jingbo Zhu, 2011). A multi-aspect segmentation model is proposed to handle multi-aspect sentences. Finally, an aspect-based opinion polling algorithm is presented in detail. (Jingbo Zhu, 2011)

In (Alsaaran, et al., 2020) paper, a summarization system for opinions in restaurant reviews is proposed. The review summary for a restaurant is composed of two parts. The first part is the percentages of positive and negative opinion words and phrases among all the opinion words and phrases that exist in the reviews of this restaurant (Athindran, et al., 2020). This gives the user an idea about the contrast between the positive and negative opinions in the restaurant’s reviews, instead of just a single number that represents the average rating. The second part of the review summary is an image that represents a word cloud for all the opinion words and phrases in the reviews of the restaurant (Lu, 2013). The size of the words depends on the occurrence frequency of the word in all the reviews. So more frequent words are larger in size. (Alsaaran, et al., 2020). The idea here is that people have different perspectives and priorities when evaluating a restaurant. Alsaaran defined in his experiment that the proposed system represents the opinion words in a word cloud so that users can see which opinions matter to them and judge the frequency of these opinions based on their size. This will be a more useful and helpful presentation to the user than just the average user rating for a restaurant. (Alsaaran, et al., 2020).

Another method for analysing texts was seen as Word-Of-Mouth (WOM), followed in the area of marketing and promotional views (Lu, 2013). WOM has a huge impact on product sales where a positive review will encourage customers to buy that product and negative review will discourage them (Lu, 2013). Dispersion is the extent to which WOM is spread across. WOM depends upon calculating the Valence and Volume (Lu, 2013). Valence is the score of evaluation of a product by a customer whereas the volume is quantity of comments/reviews on the specific product (Lu, 2013).

Customer feedback is now an inherent element of the restaurant cycle and influences the quality of their business (Alsaaran, et al., 2020). To order to analyse the mood of the comments received to restaurants, a variety of methods have been pursued

. This is basically divided into two parts:

a. Creating a Bag-Of-Words dictionary (Lu, 2013)

b. Finding the polarity and strength of sentiments and then classifying the word accordingly (Lu, 2013).

SentiWordNet is a public sentiment lexicon that’s used to extract sentiments of WOM (Hung & Lin, 2013). A word in SentiWordNet usually contains several senses where each sense has its own sentiment value in three sentiment orientations (Nicholas Cummins, 2018). The first sense for a word is picked up in its assigned part-of-speech tag in SentiWordNet, because this sense is generally the most common usage (Hung & Lin, 2013). The sentiment value is obtained, and the sentiment value of a sentence can be accumulated by words. Each synonymous set in SentiWordNet contains three sentiment tags in positive, negative, and objective orientations, respectively (Hung & Lin, 2013). For word i, posWi indicates the sentiment value in positive orientation, negWi indicates the sentiment value in negative orientation, and objWi indicates the sentiment value in objective orientation. Thus, the sum of posWi, negWi, and objWi for word i equals 1 (Hung & Lin, 2013). A word whose sentiment value is the greatest in positive, negative, or objective orientation is defined as a positive, negative, or objective word, respectively. SVM has performed effectively for classification in the literature textual reviews (Hung & Lin, 2013).

Bag of words model is used to model texts numerically using various weighting schemes such as ngram, Boolean, t.f and I.d.f. (Alshari, et al., 2018). However, Alshari demonstrates in this experiment that this model lacks the crucial step of sentiment information which in turn led to discovery of Word2Vec in both the continuous bag-of-words (CBOW) and the skip-gram (SG) models for discovering semantic of words for various Natural Language Processing tasks (Alshari, et al., 2018).

Word2vec is used to calculate the words vector, and the weighted word vector is obtained by combining them (Hao Tian, 2018). The core idea of Word2vec is to map the sentences into a higher-dimensional matrix, and replace the semantic relations between words with the mathematical relations in the matrix, so that the computer can understand the sentences in natural language through mathematics, and finally achieve the effect of making the words in similar contexts have similar vectors, Word2vec includes two models: CBOW (Continuous Bag-of-Words) and Skip-gram (Hao Tian, 2018).

## Objectives:

### 3.1 Business Problem:

With the rapid development of the Internet, there are many excellent Internet applications in the market, including social applications like yelp.com, TripAdvisor that collects customer’s feedback on businesses that they have visited (Gara, 2013). Huge amount of raw data is collected from numerous pages and stored. This raw data has no meaning or value to it. Manual process of excel, analysis of other sorts are not possible and will not produce anything meaningful. (Jamie, 2012).

Research can be qualitative, quantitative or based on an experiment. Qualitative research is basically focused to find the underlying principles of the subject whereas in the quantitative analysis the findings are converted to numerical matrix (Stephanie, 2016). Experimental data is based on a secondary dataset available through license on which a number of different experiments are conducted in order to gain some insights (Stephanie, 2016).

This is an experiment based approach where I will analyse the secondary data that was procured from yelp.com, I will be creating top positive and negative words for each restaurant in the dataset available for quick deductions regarding the best factors and areas of improvement.

### Research Question:

Analysis of the various cuisine types across restaurants in the US and to find the area of improvement and praise for restaurants.

### Aim:

Showcase the various cuisine types and highlight the negative and positive feedbacks of a restaurant, so that they can work on improving the negative aspects while maintaining the positive aspects of the restaurant. Thus, enabling restaurants to understand the consumer demands and help increase the sales, profits and service.

### Scope:

This project is developed for restaurant owners, investors or any other relevant stakeholder. The restaurant data includes reviews, ratings, and location etc. of all the restaurants in the United States. This data will be analysed to produce accurate positive and negative feedback for the top 10 restaurants of the top cuisine type.

### Limitations:

This project does not take into consideration ‘Sarcasm’ as a factor, which is a major drawback.

### 3.6 Roadmap

In order to achieve the objectives, we shall follow Figure 1: Steps to analyse data, where step by step roadmap is explained.

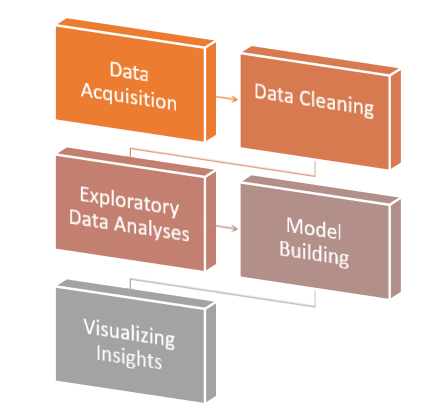


Figure 1: Steps to analyse data

## Methodology:

In this research, the data being used is secondary data, which is defined as the data already present or collected by others (Lucas, 2013). In our case, the data is provided by Yelp.com. For further analyses, we have conducted qualitative research where we analysed the dataset carefully to look out for important and meaningful visualizations and studied the new insights and trends (Udemy, 2020).

In this research, we shall follow the experiment based approach to analyse the restaurant reviews based on cuisine types using Support Vector Machine. We have numerous different cuisines out of which we have filtered 16 different types of cuisine, to figure out what factors contribute in making a good restaurant and factors which concerns customers which will further help us decide in what areas the restaurant needs improvement and growth.

In this experimental research we follow the standard framework of CRISP-DM Methodology, where data collection and data cleaning hold majority value of the project (Rüdiger Wirth, 2004). In this project, we will analyse the reviews of the users to help figure what factors relate behind a customers’ review. In this process, we also plot different restaurants according to the review counts city-wise and state-wise to understand the basic expansion of the dataset.

The figure below shows the steps included in the project:

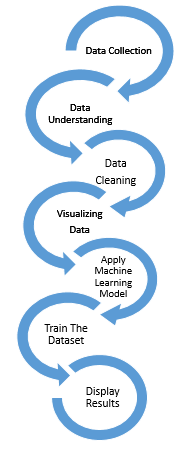


Figure 2: Project Plan

### Data Acquisition:

Yelp is a vast website that acquires reviews from users for a local business which can include restaurants, stores, retail chains and many more (Relations, 2020). Procuring yelp data was a task since we have a huge set of data available. In order to extract the data from yelp.com, web scraping method was chosen so as to extract on the required data.

Web scraping is a process of extracting the data automatically from a website (WebHarvy, 2016). This can be done using applications available such as WebHarvey, Outwit Hub, and Visual Web Ripper as well as by writing a code (Lawson, 2015) Web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet’s seemingly endless frontier (Toth, 2010-2020).

Web scraping mechanism consists of two modules namely, Web crawler and Web scraper. In general terms, crawler is the car, as the name suggests it ‘crawls’ through the web pages it moves sequentially from one page to another as a normal user would do and scraper is the driver which will control the crawler, that is, the number of pages it should go up to, the specific data that needs to be scrapped off and other such commands. (Lawson, 2015) It is an intelligent way to extract huge data without the hassle of manually copy pasting it to desired application (Lawson, 2015). To get through with the web scraping, python code was created by me using various libraries which are described below.

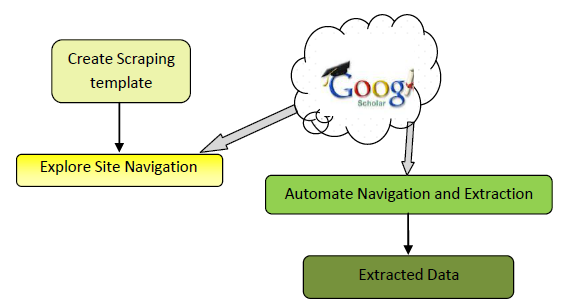


Figure 3: Block Diagram of Web Scraping

Importing [urllib.request](https://docs.python.org/3/library/urllib.request.html#module-urllib.request) module defines functions and classes which help in opening URLs (mostly HTTP) in a complex world — basic and digest authentication, redirections, cookies and more (Python, n.d.), opens up the yelp webpage from where the data needs to be extracted. Using findall() function and defining the div and class for the element we are trying t scrape the element will be loaded in our python temporary data frame. Next step is to write the temporary data frame into a csv using write operation.

Using the web scraping method, we could scrape the username, date of review, location and the review content. As shown in Figure 4: Web Scraping Output, we can see the result of one review we extracted for test run.

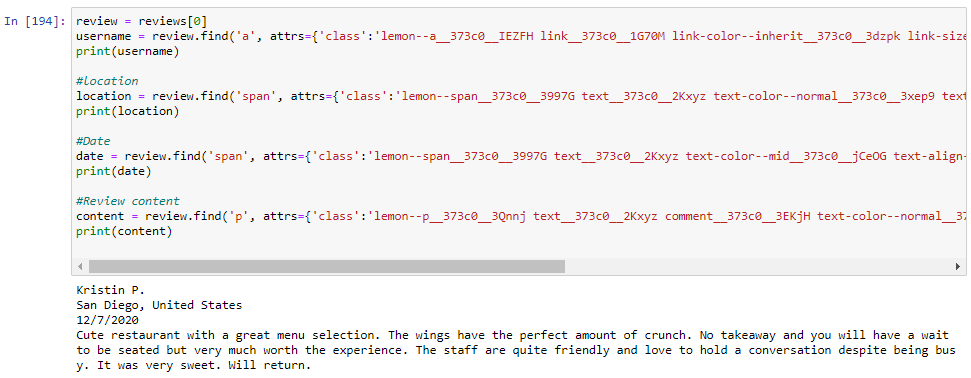


Figure 4: Web Scraping Output

Moving further, this was run through for loop for gathering all the reviews at once. Later on, extraction of top 10 restaurants serving steak in Dublin was done using the web scraper. We had 1000 records for restaurants combined. Yelp Ireland did not have enough reviews. The reviews ranged from a meagre 50 to 1000 reviews. Thus, 1000 records were too few for making any accurate deductions on restaurants also another major factor was that Yelp detected scraping from its website and blocked my computers access to it. Hence back up plan was reinstated.

### About the Dataset:

From the yelp website, open source data set was chosen where the complete dataset is available to users to explore and work on it (Yelp, 2020). It consists of one huge compressed file of 9.87 GB which when extracted leads to json files stated in the following table:

|  |  |
| --- | --- |
| File Name | File Size |
| user.json | 2370 MB |
| business.json | 131.87 MB |
| review.json | 5099.75 MB |
| tips.json | 233.20 MB |
| photo.json | 24.47 MB |

Figure 5: List of Files in Yelp Dataset

The yelp data set consists the following data:

* 192,609 businesses.
* 6,685,900 reviews.
* 1,223,094 tips.
* 1,637,138 users.
* 200,000 pictures.

As stated in the Figure 5: List of Files in Yelp Dataset, yelp data set consists of users ( unique user\_id as registered on yelp), businesses (unique business\_id as registered on yelp), Reviews written by the users, tips given, ratings given to each business by user, average rating of the business. In this research, we will consider on the Business and Review JSON files to execute our objective.

In the Business JSON file, each record consists of business\_id, categories, name, city etc which are listed in detail in the following table:

|  |  |
| --- | --- |
| Attribute Name | Description |
| business\_id | 22-character unique business id (object) |
| categories | an array of strings of business categories |
| name | Name of Business (String) |
| city | city name to where the business belongs (String) |
| state | State name 2-character (string) |
| postal code | 5 digit code of the business |
| latitude | Float, Latitude |
| longitude | Float, Longitude |
| stars | Float, average rating of the business |
| review\_count | integer, sum of number of reviews |
| attributes | An array of strings highlighting different aspects of business. |

Figure 6: Description of Business JSON

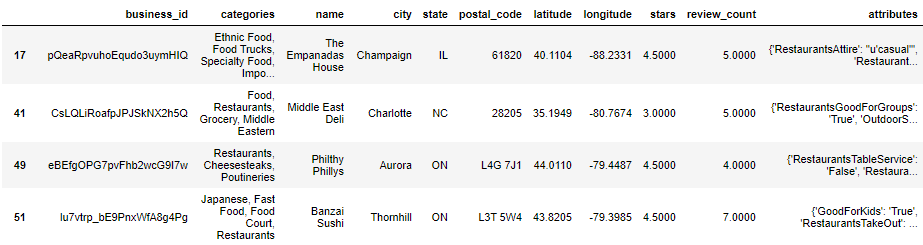


Figure 7: Business Dataset

In the Review JSON file, each record consists of unique user id, business\_id, review\_id, review of the user stated as follows:

|  |  |
| --- | --- |
| Attribute Name | Description |
| business\_id | 22-character unique business id (String) |
| user\_id | 22-character unique user id (String) |
| review\_id | 22-character unique review id (String) |
| Stars | Integer, star rating |
| useful | integer, number of useful votes received |
| funny | integer, number of funny votes received |
| cool | integer, number of cool votes received |
| date | date |
| text | review of the user (string) |

Figure 8: Description of Review JSON

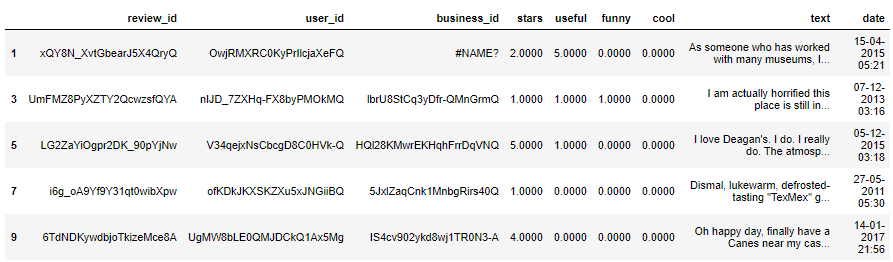


Figure 9: Review Dataset

Each record in photo json include photo\_id, business\_id described in Figure 10: Description of Photo JSON:

|  |  |
| --- | --- |
| Attribute Name | Description |
| photo\_id | 22-character unique photo id (string) |
| business\_id | 22-character unique business id (string) |
| Label | category to which photo belongs (string) |

Figure 10: Description of Photo JSON

Each record in the tip JSON gives details of the user\_id, business\_id and count of the compliments received.

|  |  |
| --- | --- |
| Attribute Name | Description |
| user\_id | string, 22-character unique user id |
| business\_id | string, 22-character unique business id |
| Compliment Count | integer, number of compliment counts |

Figure 11: Description of the Tip JSON

Each record in the user JSON gives details of the user\_id, review count, average stars, date since the user started to yelp, etc.

|  |  |
| --- | --- |
| Attribute Name | Description |
| user\_id | string, 22-character unique user id |
| name | String, first name of the user |
| review count | Number of reviews given by the user |
| yelping since | Date, date since the user started to yelp |
| friends | array of string, list of user’s friends |
| useful | integer, number of useful votes by the user |
| funny | integer, number of funny votes by the user |
| cool | integer, number of cool votes by the user |
| fans | integer, number of fans of the user |
| elite | date, year since user became elite member |
| stars | average rating of all reviews |
| compliment\_hot | integer, number of hot compliments received by the user |
| compliment\_more | integer, number of more compliments received by the user |
| compliment\_profile | integer, number of profile compliments received by the user |
| compliment\_cute | integer, number of cute compliments received by the user |
| compliment\_list | integer, number of list compliments received by the user |
| compliment\_note | integer, number of note compliments received by the user |
| compliment\_plain | integer, number of plain compliments received by the user |
| compliment\_cool | integer, number of cool compliments received by the user |
| compliment\_funny | integer, number of funny compliments received by the user |
| compliment\_writer | integer, number of writer compliments received by the user |
| compliment\_photos | integer, number of photos compliments received by the user |

Figure 12: Description of the User JSON

### Data Cleaning:

This is the data pre-processing step followed by data acquisition from various places. In this case study, we have acquired data from yelp and then converted the large JSON files to csv files by writing a python code to parse the data through it.

Reading the data from csv using read\_csv it was noticed that extra blank lines were inserted in the csv shown in Figure 13: Raw Business Data. For better analyses and code, getting rid of them was mandatory.

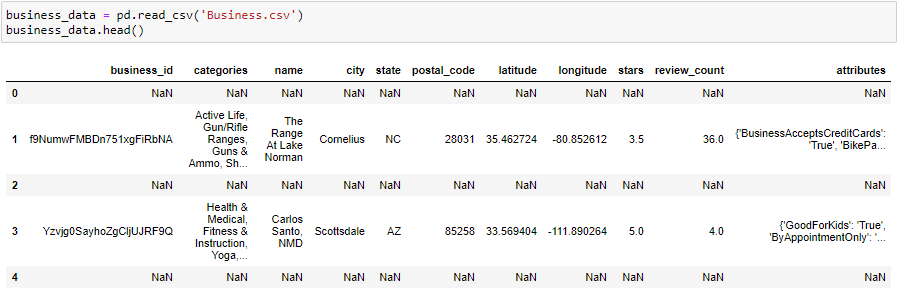


Figure 13: Raw Business Data

This goal was achieved by dropping the null values throughout the csv file. Moving further, this case study was related to evaluating the restaurants based in the United States. Hence, data was filtered by allowing all 50 states of US as shown in Figure 14: Data Filtering.

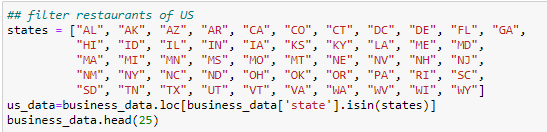


Figure 14: Data Filtering

In this research, from the business dataset we want to pick up only restaurants hence we selected column categories which consisted of all local businesses and restricted only those values that consisted of the word restaurant in them as shown in Figure 15: Restaurant Selection.



Figure 15: Restaurant Selection

For further analyses, creation of new column in the data frame was done labelled as categories. This column consisted of all different cuisines available in the dataset after filtering for restaurants.

The following category list was made with the following cuisine types:

1. American
2. Mexican
3. Italian
4. Japanese
5. Chinese
6. Thai
7. Mediterranean
8. French
9. Vietnamese
10. Greek
11. Indian
12. Korean
13. Hawaiian
14. African
15. Spanish
16. Middle-eastern

This is shown in the Figure 16: Cleaned Business Dataset

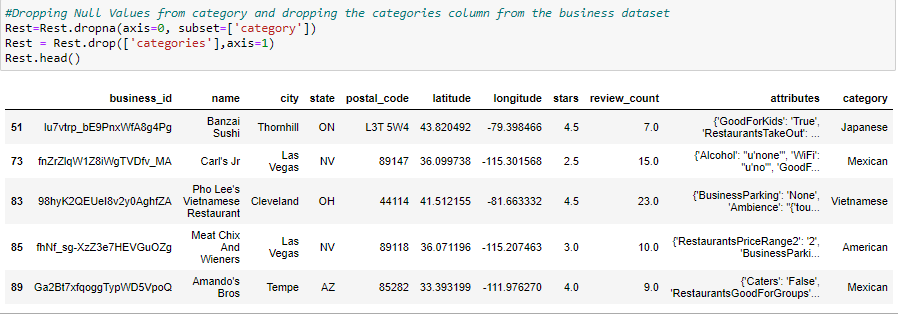


Figure 16: Cleaned Business Dataset

Following the roadmap, further review dataset is loaded through the Jupyter notebook using read\_csv() function.

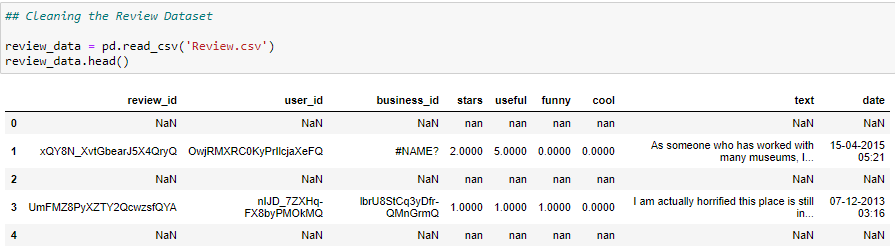


Figure 17: Unclean Review Dataset

Similar to business dataset, we see that multiple null values have been loaded in the csv. Hence, dropping these null values is the next step followed by count of total reviews available for analysis.

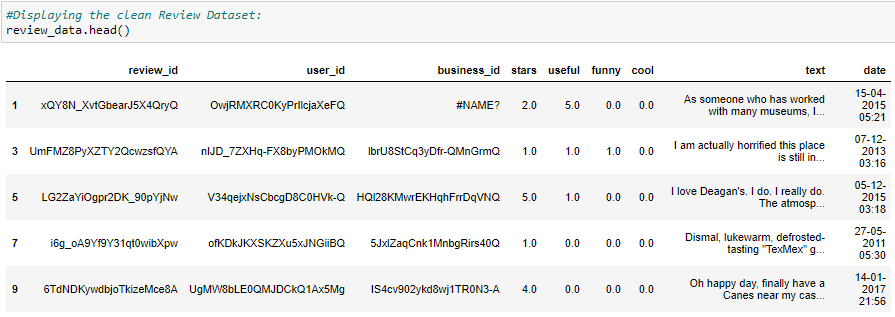


Figure 18: Cleaned Review Dataset

Merging business and review data for further analysis was done by using inner join based on the basis of business\_id and the result were as follows:

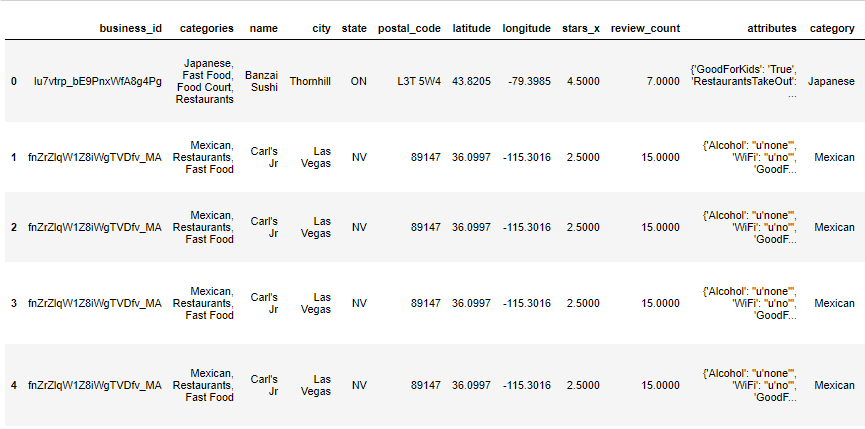


Figure 19: Merged Data Part 1



Figure 20: Merged Data Part 2

As you can see in the Figure 19: Merged Data Part 1, Carl’s Jr has appeared consecutively 4 times in a row but as we move further through the columns in the merged dataset we see that these are different reviews as they as different review id as seen in Figure 20: Merged Data Part 2. In the final merged dataset, a total of 71635 records were found.

Following this step, renaming of columns such as star\_x (originally from the business data set) to avg\_star and star\_y (originally from the review dataset) to review\_star was done for better understanding. Check for duplicate rows were made to ensure we do not have and duplicated data in our dataset.

For exploring the data, we removed all the special characters from the Reviews text columns such as comma, quotation marks, question mark, exclamation mark etc. For easy and better analyses, we have created a column named labels based on the review\_star column present in the merged dataset. For all the review\_star values greater than 3 we have labelled them as positive and for all the review\_star less than 3 we have labelled them as negative. For the review\_star values equal to 3 we have kept them as neural since according to our analyses they are middle values and won’t serve much purpose for data analyses.

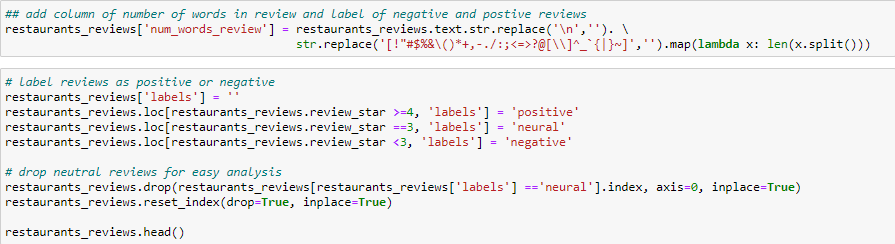


Figure 21: Creating Labels column and removing special characters from review text

## Exploratory Data Analysis (EDA):

EDA is process to understand the data by summarizing the important characteristics of dataset and plotting them visually (Prabhu, 2019). Plotting helps to understand the numbers of different characteristics in your dataset which is an important step for modelling your data since you understand the parameters of your train set more efficiently. (Prabhu, 2019).

In this study, ggplot library is used to plot all the graphs (Brownley, 2016). Ggplot is a plotting system in python which is very expressive and has a very appreciative API (Brownley, 2016). It is based on R’s ggplot2 and deals with the grammar of graphics. With the help of ggplot, we can reuse the same code to create different types of visualizations and hence spend less time in coding whereas more time interpreting what the graph is saying and where is it leading to us with the current dataset (Brownley, 2016). Many different types of graphs can be created such as bar graph, heatmap, scatterplot and many more. In Figure 22: Restaurants by Category, we can see that American cuisine has highest number of restaurants in the dataset followed by Mexican and Italian. The least number of restaurants are for Spanish and African cuisine. We have analysed the reviews for most number of restaurants available which is American here.

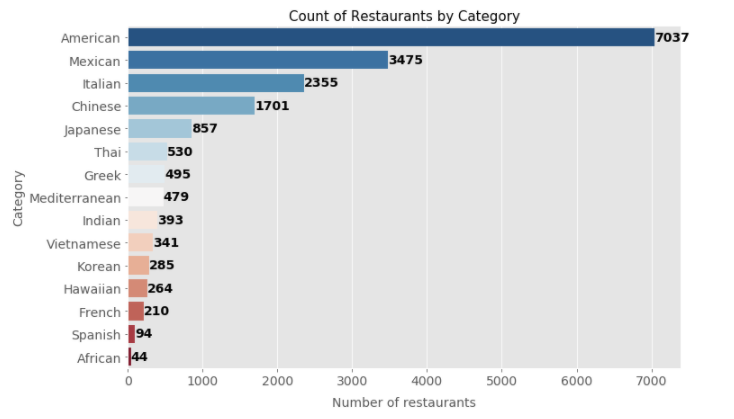


Figure 22: Restaurants by Category

In Figure 23: Restaurant by State, we have combined the state value counts according to the business\_id present in that state and visualized it using the bar plot. We infer that state of Arizona and Nevada have most of the restaurants as compared to all other states.

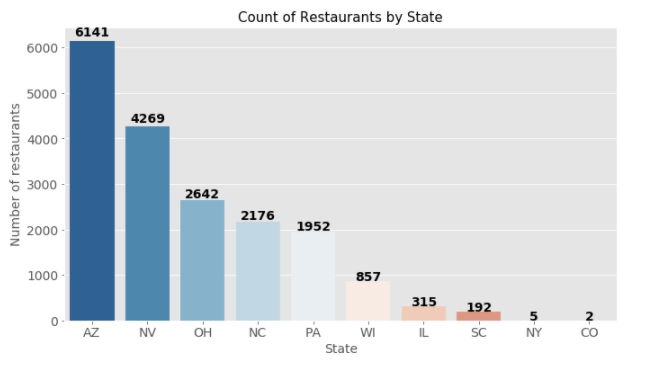


Figure 23: Restaurant by State

As shown in Figure 24: Restaurants by City, we can see that Las Vegas and Phoenix has the highest number of restaurants. In comparison with Figure 23: Restaurant by State, Arizona is the highest state with number of restaurants whereas Phoenix, capital of Arizona, has second highest number of restaurants and Nevada, having second highest ranking in the number of restaurants, Las Vegas has score of highest number of restaurants.

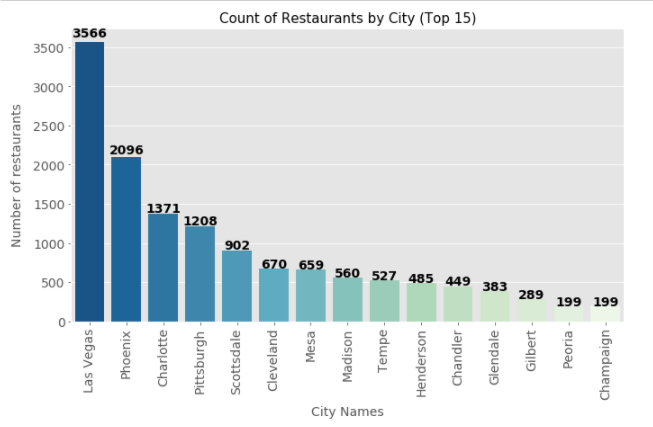


Figure 24: Restaurants by City

Figure 25: Reviews by Cuisine, depicts the sum of number of review counts for each cuisine. American cuisine has the highest number of reviews whereas African cuisine has the lowest number of reviews. This is also verified by the fact that African restaurants have the least number as seen in Figure 22: Restaurants by Category

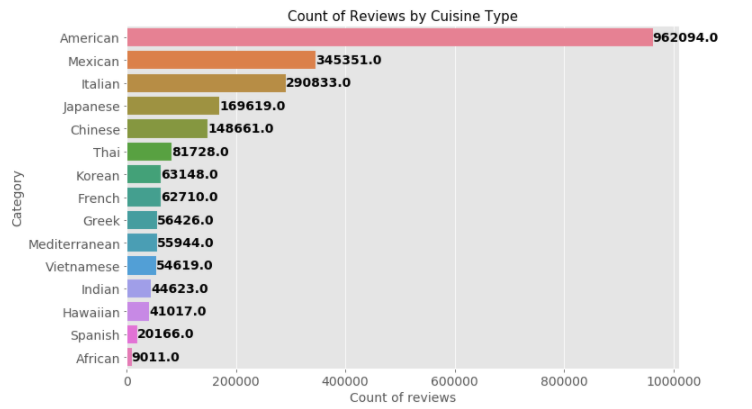


Figure 25: Reviews by Cuisine

From Figure 26: Top 5 States with most reviews, the most reviews were received for state of Nevada which has second highest number of restaurants.

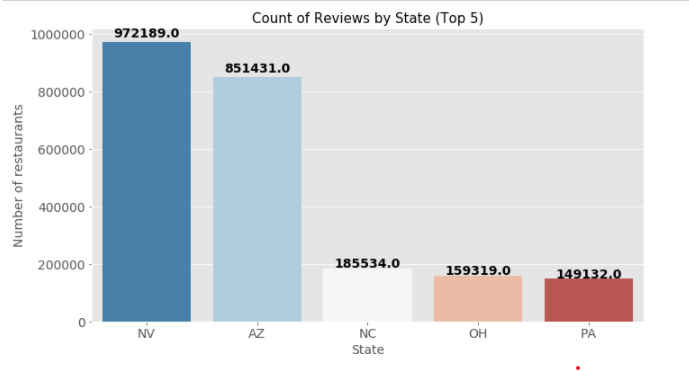


Figure 26: Top 5 States with most reviews

In reference to state wise top 5 review, Figure 27: Top 10 cities with Most Reviews shows that Las Vegas has the highest number of reviews.

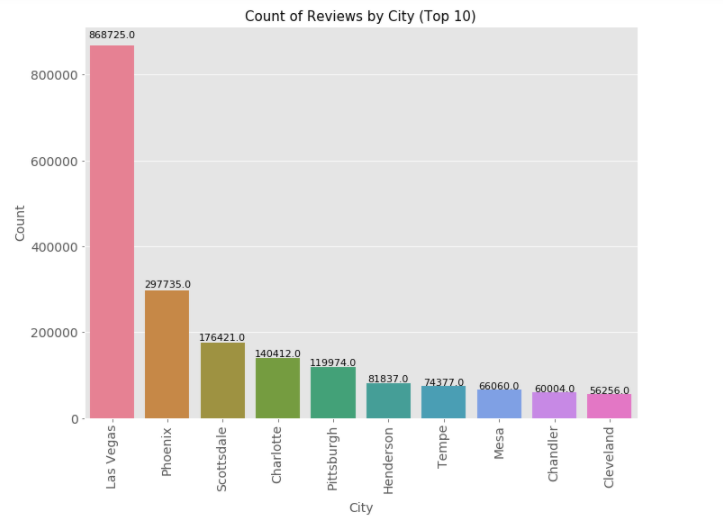


Figure 27: Top 10 cities with Most Reviews

Figure 28: Top 10 Restaurants with most reviews shows top 0 restaurants with most reviews present overall in that dataset.

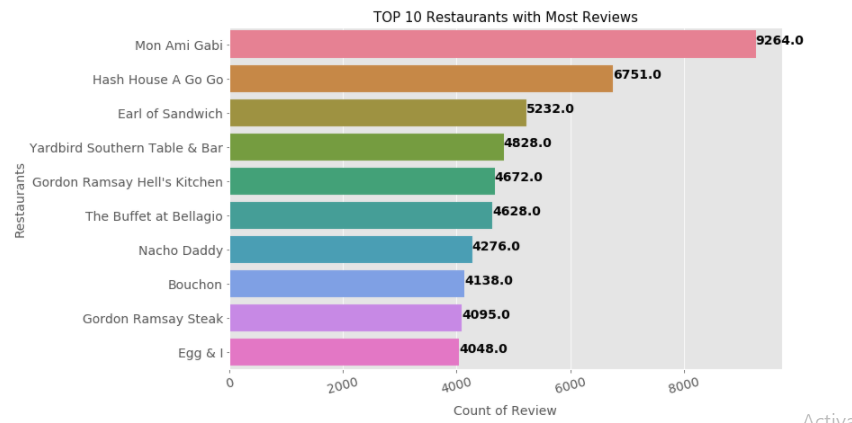


Figure 28: Top 10 Restaurants with most reviews

Figure 29: Top 10 restaurants with American Cuisine gives out to be analysed dataset. We will be analysing the top 10 restaurant reviews based on American cuisine since it holds highest number of reviews.



Figure 29: Top 10 restaurants with American Cuisine

Figure 30: Count of Restaurants against Ratings defines the distribution of ratings of all the restaurants available in our dataset.

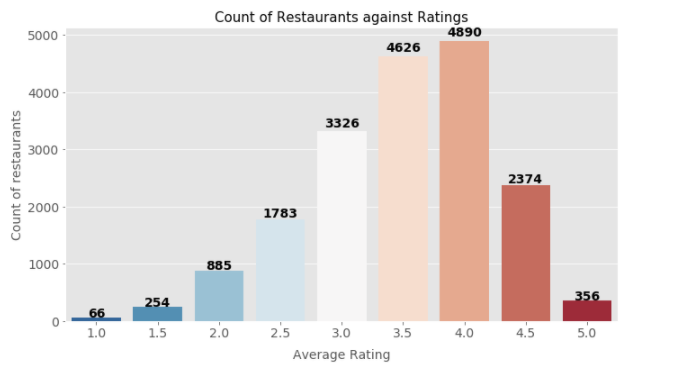


Figure 30: Count of Restaurants against Ratings

Figure 31: Percentage of Positive Reviews by Cuisine type, shows that American cuisine has 75% positive reviews whereas Greek and Korean have the highest percentage of positive reviews despite having 56426 and 63128 reviews respectively.

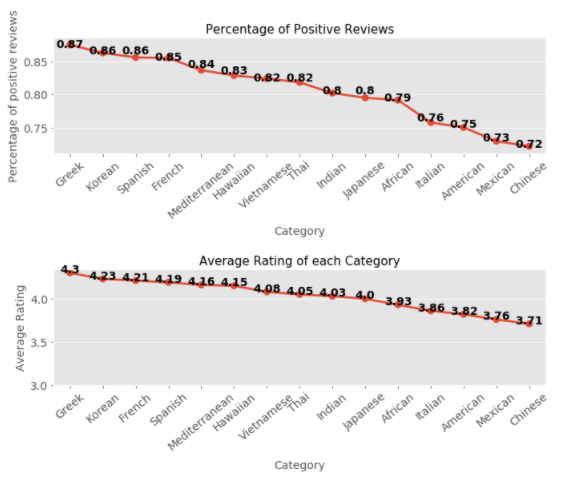


Figure 31: Percentage of Positive Reviews by Cuisine type

Figure 32: Average rating of each category : Average rating of American restaurant is 3.82 despite having the most number of reviews. Hence it can be concluded that more than half of the customers were happy eating at such cuisine types.

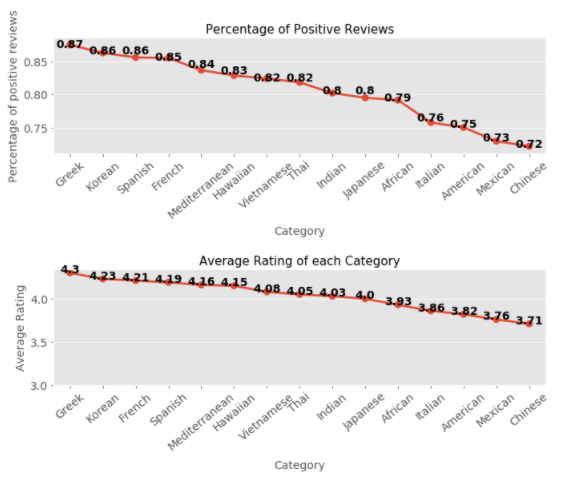
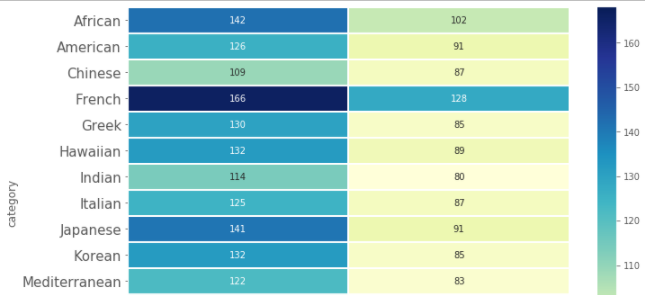


Figure 32: Average rating of each category

Figure 33: Heat map of positive and negative words in Review, shows the average count of the positive and negative labels in each cuisine type.



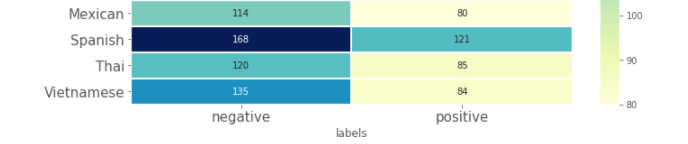


Figure 33: Heat map of positive and negative words in Review

Figure 34: Distribution of Review Stars, shows how the users have rated all the restaurants in our dataset. In order to sample our dataset, we will provide dummy ratings records for balanced dataset.

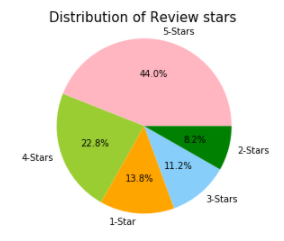


Figure 34: Distribution of Review Stars

## Modelling the Data:

The building blocks of a successful machine learning algorithm is the training dataset (Keith Norambuena, 2019). The training dataset is fed to an algorithm which in return generates new set of rules based on the conclusions driven from the dataset (Brownley, 2016). By feeding different data several times the machine learning model can be made more accurate (Brownley, 2016). Surmising inferences from the training dataset is the core strength of the machine learning algorithm (Brownley, 2016).

Therefore, as per our research, we are going to go through with Support Vector Machine in this dataset, which is recognised in supervised machine Learning. The process of modelling the data is where you train the data, called as labelled data, according to features, predictor variables which are used to train the model to predict the labels (Bishop, 2006). Then, they are tuned according to the business needs and visualized in graph.

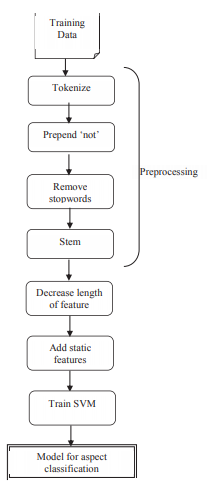


Figure 35: Model Data in SVM (Bishop, 2006)

In this experiment for the process of modelling, conversion of textual data in the reviews column is done to lower case for similarity in all words followed by removing the punctuation marks from the same.

Two files were downloaded from the internet that consists of 2041 Positive and 4818 negative words that were found to be unnecessary and meaningless. They were loaded into the code as positive\_words and negative\_words respectively. (Lexicon, n.d.)

Support Vector Machines is a classification algorithm that works on Supervised Machine learning (M.N. Murty, 2016). It categorizes a set of labels into a new category and then plots it (M.N. Murty, 2016). We are going to use scikit library to implement the support vector machine algorithm (M.N. Murty, 2016). In the next step, American restaurants were selected in the dataset and split into the train set and test set with a test size of 0.5.

Using CountVectorizer(), we convert the review text documents into a token of words. This implementation produces a sparse representation of the counts using scipy.sparse.csr\_matrix. This method boils down to just counting how many times every word appears in a text and dividing it by the total number of words.

Later, we run the model and get Linear SVC output as follows:

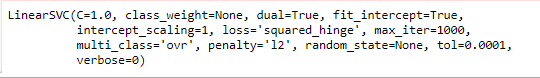


Figure 36: Output of SVC Model

C : Regularization Parameter

Class\_weight : Set the parameter C of class i to class\_weight[i]\*C for SVC. If not given, all classes are supposed to have weight one. The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n\_samples / (n\_classes \* np.bincount(y)) (python, 2020).

Dual: training set performs dual or primal optimization.

Fit\_intercept: whether intercept needs to be calculated.

Intercept\_scaling: When self.fit\_intercept is True, instance vector x becomes [x, self.intercept\_scaling], i.e. a “synthetic” feature with constant value equals to intercept\_scaling is appended to the instance vector. The intercept becomes intercept\_scaling \* synthetic feature weight

Loss: specifies the loss function

Max\_iter: maximum iterations that needs to be performed.

Multi\_class: defines the class strategy for the model

Penalty: specifies the norm used in the penalization.

Tol: tolerance for stopping criteria

For the next step, we store the score of words calculated from the review text in a temporary database and then count the frequency of each word in the text. To calculate the polarity we follow the following steps:

* 1. Calculate score of each word
  2. Calculate frequency of each word
  3. Dimension of the data frame formed

Following that we store the top 10 positive and negative words in an array and then display them on a graph to see and analyse what a customer likes and dislikes the most at any given restaurants under American Cuisine. This is shown in Figure 37: Words in American Restaurant

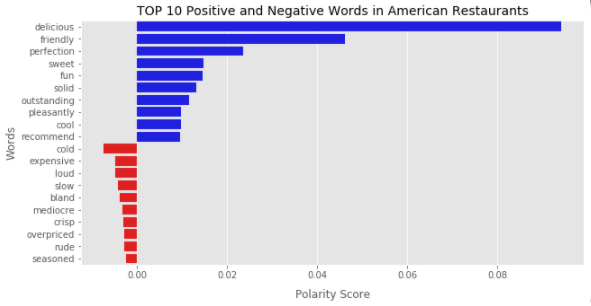


Figure 37: Words in American Restaurant

We apply SVM to predict the stars rating in each cuisine type. We use TfidVectorizer to break the texts into single words and bi-grams. It then calculates TF-IDF representation (M.N. Murty, 2016). Applying the fit method we build the library and transform converts the texts into numerical matrix.

We then predict the star ratings and compare them to actual star ratings.

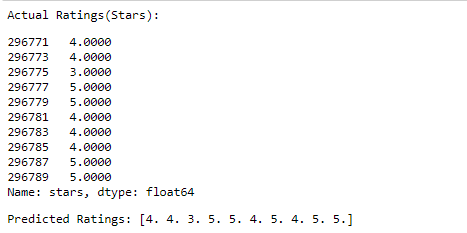


Figure 38: Actual VS Predicted Ratings

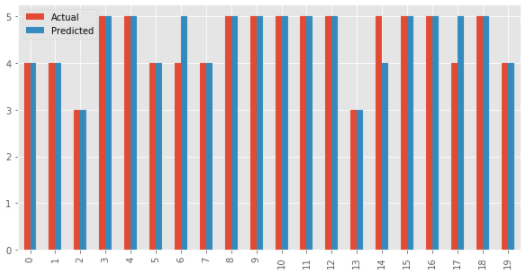


Figure 39: Graph of Predicted VS Actual Ratings

As we can see from the Figure 38: Actual VS Predicted Ratings and Figure 40: Classification Report the accuracy is 67% and precision is 70%.

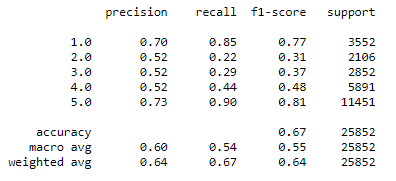
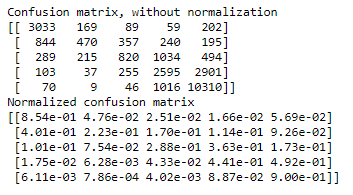
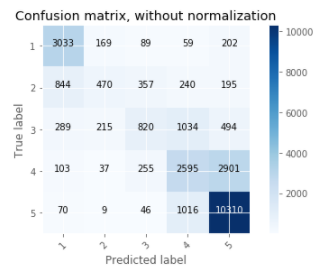


Figure 40: Classification Report





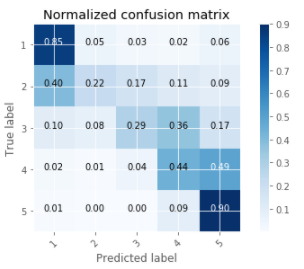


Figure 41: Confusion Matrix

## Results and Findings:

These are the predicted ratings for Top 10 American cuisine based restaurants

### The Arrogant Butcher:

This restaurant has most number of reviews in the American cuisine dataset.

As seen in Figure 42: Top 15 positive words for The Arrogant Butcher and Figure 43: Top 15 negative words for The Arrogant Butcher, friendly has a polarity of 0.04 and is recorded as the highest positive word whereas cold is recorded as top negative word with polarity of -0.006.



Figure 42: Top 15 positive words for The Arrogant Butcher



Figure 43: Top 15 negative words for The Arrogant Butcher

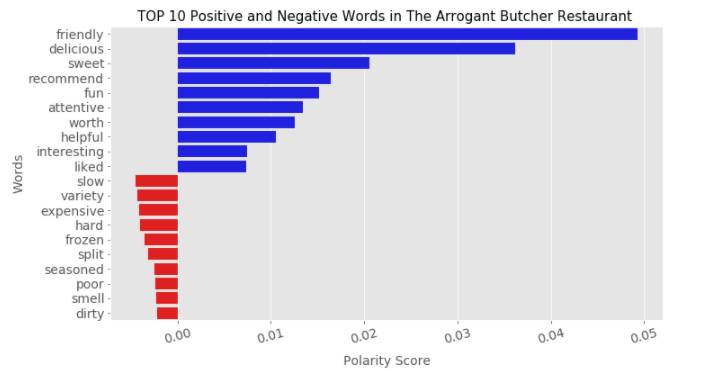


Figure 44: Graph for Top Words for Restaurant 1

As Seen in Figure 44: Graph for Top Words for Restaurant 1, “The Arrogant Butcher”: Customer service is ranked at the highest as friendly and attentive staff whereas the food is of the second best attribute as delicious food. The customers have found this place to be worth of its prices. In the areas to be worked upon, we can see that cold ranks highest. We can infer that even though the food was delicious many customers found it cold. Also, there were instances where wrong food was being served to the customer.

### Mastro's Ocean Club

In Figure 45: Top Positive Words for Maestro’s Ocean Club we note that frequency of worth is 112 but polarity is 0.0125 whereas frequency of fun is 86 and polarity is 0.015.

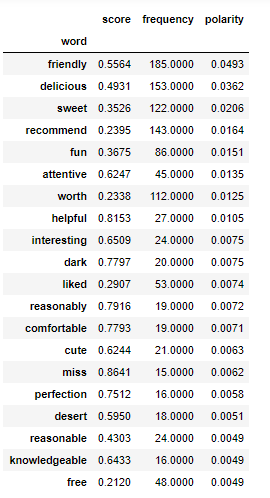


Figure 45: Top Positive Words for Maestro’s Ocean Club

 (Yelp, 2004)

Figure 46: Top Negative Words for the Maestro's Ocean Club

From Figure 47: Graph of Top words for Restaurant 2, we note the top positive and negative words for Restaurant 2.

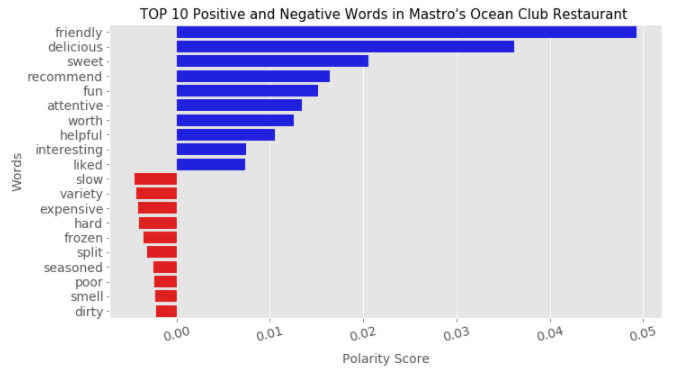


Figure 47: Graph of Top words for Restaurant 2

“Maestro’s Ocean Club”: For Customer reviews we have attentive and friendly staff which can really boost up a customer’s mood. However, they need to focus on their food area since many customers have complaint regarding the food quality and food prices.

### Strip steak

Figure 48: Top Positive Words for StripSteak show that the word “outstanding” has appeared as a distinctive feature as compared to previous 2 restaurants. Figure 49: Top Negative Words for StripSteak shows that expensve has ranked as 3rd negative word.

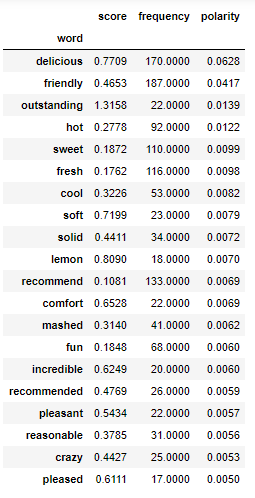


Figure 48: Top Positive Words for StripSteak



Figure 49: Top Negative Words for StripSteak

“Strip Steak”: We have got reviews to upscale the restaurant hence in order to gain more profit restaurant can consider upscaling the environment. Also, the place could be kept a bit tidy for better reviews and recommendations. For food service, it seems that the customers have really liked the soft and fresh food served to them.

Figure 50: Top Words for Restaurant 3, we can see tht comfort is ranked as the least positive word and upscaling of restauran has been suggested by many customers.

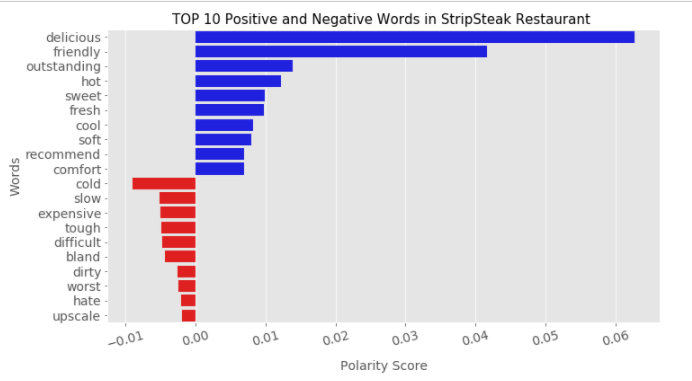


Figure 50: Top Words for Restaurant 3

### Windsor

Figure 52: Top Negative Words for Windsor, it can been seen that bland as well as seasoned food were one of the concerns of customers

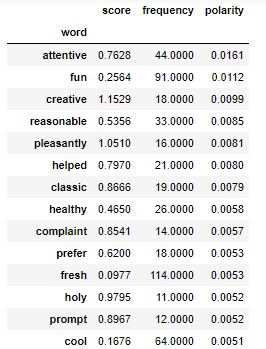


Figure 51: Top Positive Words for Windsor

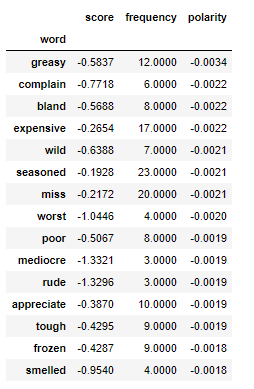


Figure 52: Top Negative Words for Windsor

Figure 53: Top Words for Restaurant 4 includes the positive comments such as fun, creative, health and reasonable whereas negative comments are expensive, poor and smelly.

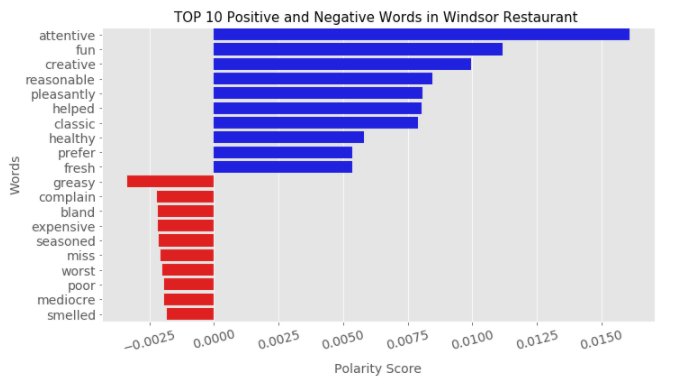


Figure 53: Top Words for Restaurant 4

“Windsor”: Customer have found this place to be interesting and fun which tells us that may be the environment infrastructure amazes them to an extreme level. Also, they find the staff to be attentive but however they get cold food or sometimes there is a missing item in their order. The restaurant needs to focus on this area.

### Lawry’s The prime rib

Figure 54: Top positive words for Lawry's and Figure 55: Top Negative Words for Lawry's depict that some customers found the staff to be friendly but the service was slow.



Figure 54: Top positive words for Lawry's

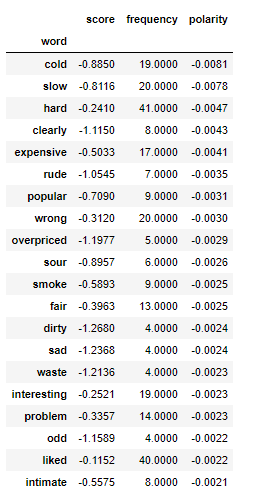


Figure 55: Top Negative Words for Lawry's

Figure 56: Graph of Top Words for Restaurant shows that few customers did not receive a great service as they found the staff to be rude.

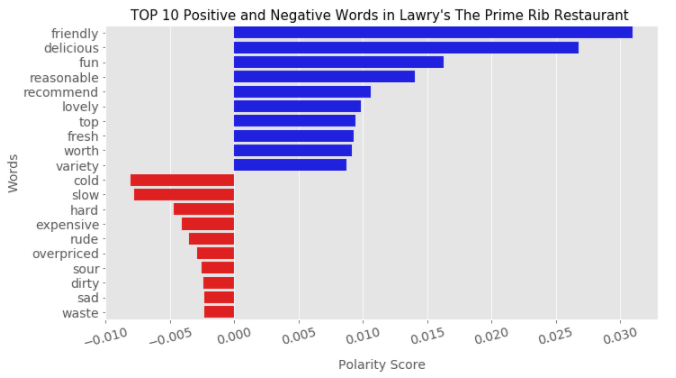


Figure 56: Graph of Top Words for Restaurant

“Lawry’s The Prime Rib”: The customers have also complained about the service being offered as rude and the price of the restaurant to be expensive or more than what it deserves to be. If the restaurant manages to work on these aspects then it can have a huge profit gain and better recommendations.

### Blue martini lounge

Figure 57: Top Positive Words for Blue Martini Lounge and Figure 58: Top Negative Words for Blue Martini Lounge indicates that some of the customers faced problems with their food as the frequency is low whereas on the other side other customers found the food fresh.

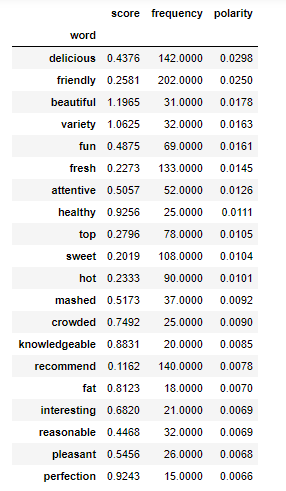


Figure 57: Top Positive Words for Blue Martini Lounge

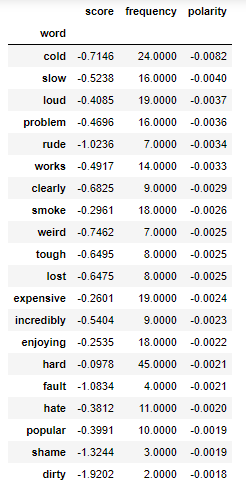


Figure 58: Top Negative Words for Blue Martini Lounge

Figure 59: Graph of top words for Restaurant 6 shows graphical representation of top positive and negatve words for Blue Martini Lounge.

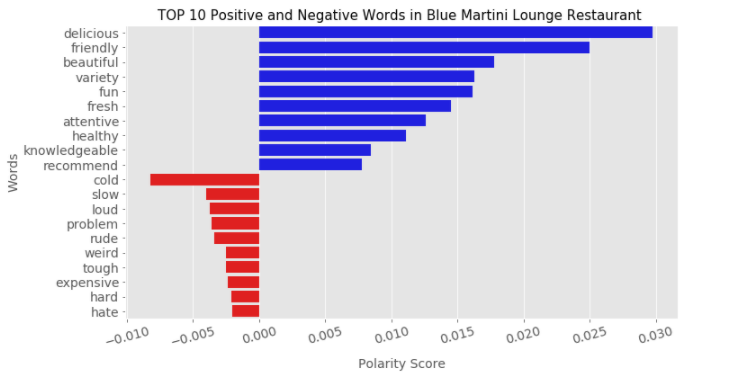


Figure 59: Graph of top words for Restaurant 6

“Blue Martini Lounge”: The key aspect in this restaurant was observed was food quality, the customers found the food to be sour although there was a variety of food available as well as the environment to be dirty and the ambience to be too loud. At the same time, the customer service was found to be helpful and attentive.

### Bahama Breeze

Figure 60: Top Positive words for Bahama Breeze and Figure 61: Top Negative words for Bahama Breeze shows that the customers enjoyed the beautiful ambience and fresh food

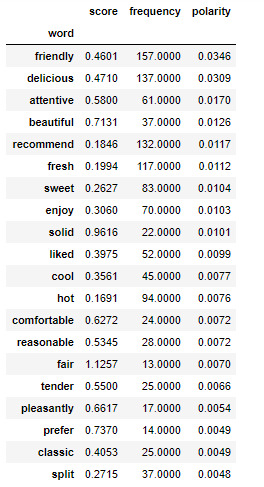


Figure 60: Top Positive words for Bahama Breeze

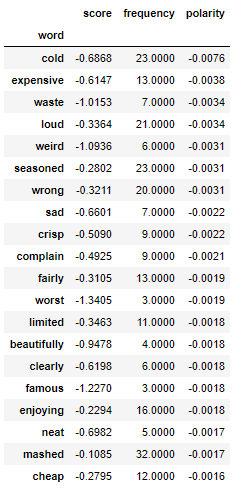


Figure 61: Top Negative words for Bahama Breeze

Figure 62: Top Words for Restaurant 7: We observe that the polarity of negative comments is much less as compared to polarity of positive comments.

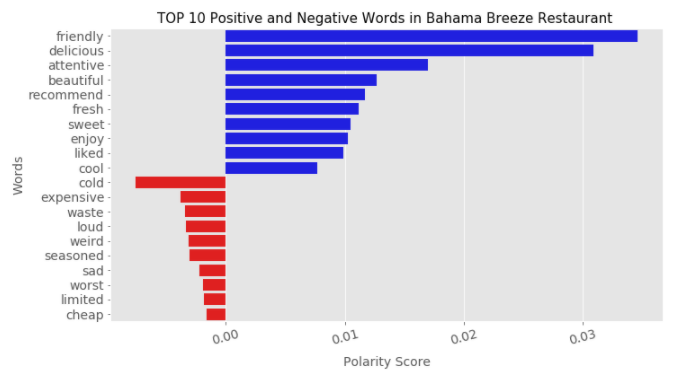


Figure 62: Top Words for Restaurant 7

“Bahama Breeze”: The customers enjoyed the beautiful view from the restaurant and found the food to be fresh and delicious. They like to recommend this restaurant for its ambience and fresh food served. For the areas of improvement, customers have commented about limited availability and seasoned food. The restaurant can apply for more seating space.

### Beckett’s Table

Figure 63: Top Positive Words for Beckett's Table and Figure 64: Top Negative Words for Beckett's Table show the review of few customers who found the food to be expensive but delicious.



Figure 63: Top Positive Words for Beckett's Table

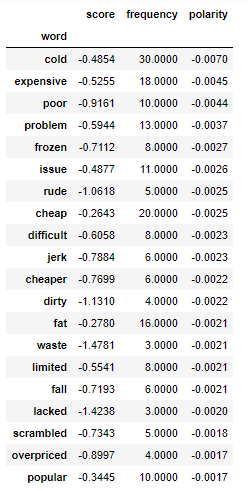


Figure 64: Top Negative Words for Beckett's Table

Figure 65: Graph for Top Words for Restaurant shows that the positive reviews were much greater in weight as comared to negative ones.

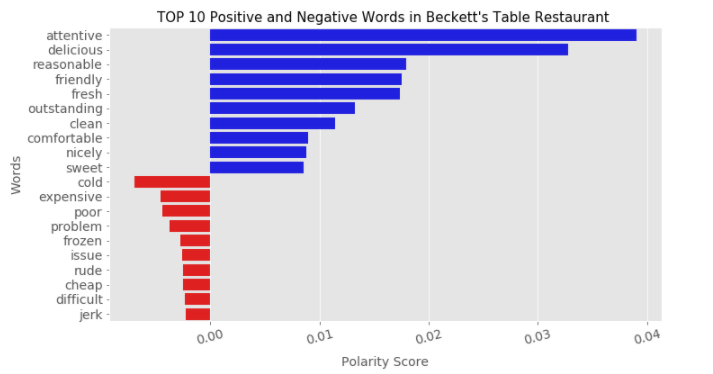


Figure 65: Graph for Top Words for Restaurant

“Beckett’s Table”: The restaurant has been rated more on the negative reviews which clearly indicates that the areas of improvement as more. The customers have found this restaurant ambience as worst, poor and cheap.

### Egg Works

Figure 66 shows that customers recommended this restaurant 125 times and polarity of the same is 0.015.



Figure 66: Top Positive Words for Egg Works

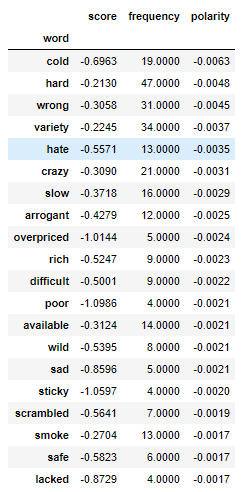


Figure 67: Top Negative Words for Egg Works

Figure 68 shows that few customers found the service to be poor nd the staff to be arrogant

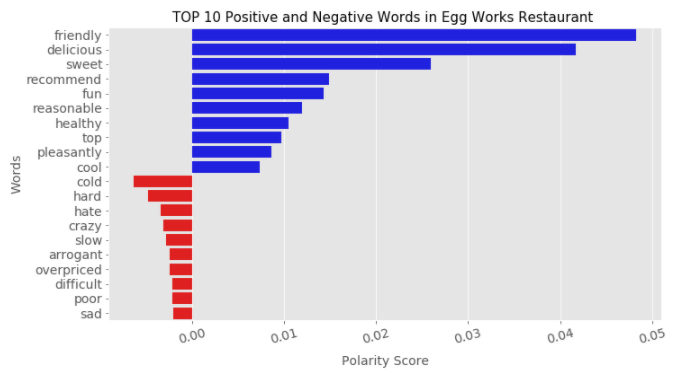


Figure 68: Graph of Top words for Restaurant 9

“Egg Works”: This restaurant has been rated excellently for their customer service and environment as clean and comfortable. Also at the same time few customers found it slow and overpriced for the portion of food served by the restaurant.

### Worth Takeaway

Figure 69 and Figure 70 denote that fresh word has the frequency of 132 but polarity of 0.02 which is lower than sweet which has a frequency of 97 but polarity of 0.03



Figure 69: Top Positive Words for Worth Takeaway

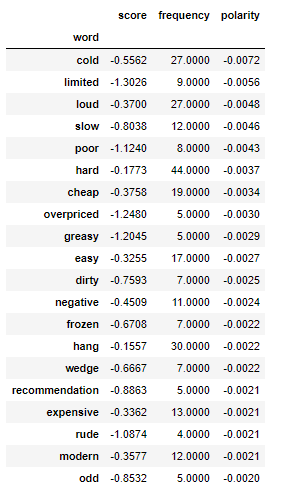


Figure 70: Top Negative Words for Worth Takeaway

Figure 71 shows the graph of top 10 words for Worth Takeaway as we can see customers loved the ustomer service of this restaurant based on words like friendly, sweet, fun, pleasant and cool.

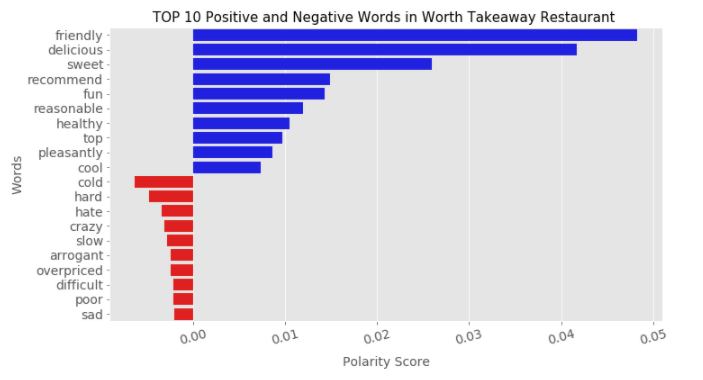


Figure 71: Graph of top words for restaurant 10

“Worth Takeaway”: This restaurant has a good customer service which is inferenced from the comments such as fast, attentive and sweet. As all restaurants have some areas of improvement this one needs to focus on the seating as well as the loud ambience.

### Combined Analysis of American Restaurants:

Figure 72: Matrix of Top Positive Words for American Restaurants consists of a table of all the positive words that have the most polarity in our model. On the basis of these words, we analyse what customers like the most about the restaurant.

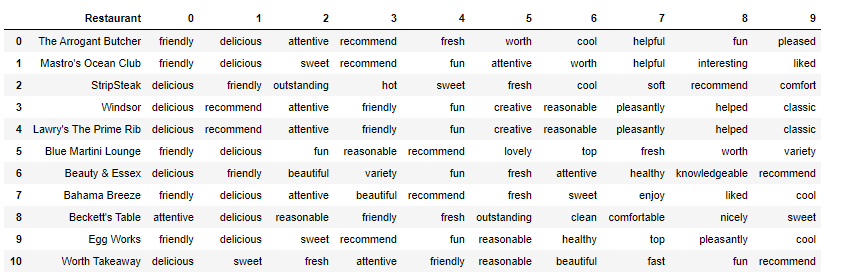


Figure 72: Matrix of Top Positive Words for American Restaurants

Figure 73: Matrix of Negative Words for American Restaurants comprises of top negative words in the top 10 restaurants of American Cuisine. Through this we can measure the areas of improvement in each restaurants.



Figure 73: Matrix of Negative Words for American Restaurants

Figure 74: Top 10 Negative Words in American Cuisine and Figure 75: Top 10 Positive words for American Cuisine shows the frequency of the words.



Figure 74: Top 10 Negative Words in American Cuisine



Figure 75: Top 10 Positive words for American Cuisine

In our proposed methodology, we found that the restaurants are highly ranked positive in terms of friendly customer services and food consumptions. Based on this they are recommended ahead by customers in their feedback. However, the food quality is a measure that defines the areas of improvement in many restaurants since the food served was found colder than usual.

In general, we will compare the food quality, customer service and prices of the restaurant. As seen in Figure 37: Words in American Restaurant, we can see that for American cuisine food taste has the highest positive polarity, meaning they most people visit these restaurants because of how delicious the food tastes. The second best attribute is supposed to be the customer service provided by these restaurants which has a polarity of 0.08. For the negative reviews, food quality is seen to be of utmost importance to customers since most of them reviewed about the food being served cold. The price seems to be expensive to many customers and the service seems to be slower than usual.

For many restaurants, we observed that the food served was bland which means that the customer was expecting more of spices or taste in their food. In few restaurants such as Maestro’s Ocean Club, Lawry’s the Prime Rib, Windsor and Bahama Breeze the food was found to be seasoned, meaning the users were looking for more availability of food throughout the year. For only 3 restaurants out of 10, the customer found their money to be worth visiting for such as The Arrogant Butcher, Windsor and The Blue Martini.

As we can see that our analyses helped us to retrieve the top words using SVM we can pass tis to the restaurants in order for them to realize what a customer likes or dislikes the most about their restaurant and what can be done in order to make their restaurant a worthwhile place to spend time upon. It is seen that customer service and food contribute to great reviews whereas the same factors can also lead to depreciation in the status of the restaurants. Hence these crucial aspects must be taken extreme care of.

Finally, we have predicted the ratings of the restaurants for American cuisine where we have received 67% accuracy and precision of 64% along with recall of 66%.

## Future Scope

Some limitations of the project include accuracy and less diversity of categorization. The training set can never be perfect as it is not an easy task to separate 18,650 reviews into good or bad just by reading them. The quality of the training data set and testing data set is a major aspect in determining the efficiency of any such algorithms. Moreover, the classification is binary and gives an overall result of the goodness or badness of the review. It does not detect sarcasm and does not deal with individual traits such as food, ambiance, cost, service among others. The future scope is to rate business based on different features, to use better and more dataset to train and to detect sarcasm

# Bibliography

Alsaaran, H. et al., 2020. Where to Eat? An Automatic Summarization System for Opinions in Restaurant Reviews. 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS).

Alshari, E. M. et al., 2018. Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2Vec for Sentiment Analysis.

Amasyali, M. F., Tasköprü, H. & Çaliskan, K., 2018. Words, Meanings, Characters in Sentiment Analysis. 2018 Innovations in Intelligent Systems and Applications Conference (ASYU).

Athindran, N. S., Manikandaraj, S. & Kamaleshwar, R., 2020. Comparative Analysis of Customer Sentiments on Competing Brands using Hybrid Model Approach. 2018 3rd International Conference on Inventive Computation Technologies (ICICT).

Aye, Y. M. A. S. S., 2017. Sentiment analysis for reviews of restaurants in Myanmar text. 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). In: s.l.:s.n.

Brownley, C., 2016. Foundations for Analytics with Python: From Non-Programmer to Hacker. In: s.l.:s.n.

Doan, T. K. J., 2016. Sentiment Analysis of Restaurant. - Anahei15th IEEE International Conference on Machine Learning and Applications (ICMLA) - Anaheim, CA, USA (2016.12.18-2016.12.20)] 2016 .

Gara, T., 2013. Fake Reviews are everyhere, how Can we catch them. [Online]   
Available at: https://blogs.wsj.com/corporate-intelligence/2013/09/24/fake-reviews-are-everywhere-how-can-we-catch-them/

Hanhoon Kang, S. J. Y. D. H., 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews.

Hao Tian, L. W., 2018. Microblog Emotional Analysis Based on TF-IWF Weighted Word2vec Model. 2018 IEEE.

Hung, C. & Lin, H.-K., 2013. Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification. IEEE Intelligent Systems ( Volume: 28 , Issue: 2 , March-April 2013 ).

Ian Goodfellow, Y. B., 2016. Deep Learning (Adaptive Computation and Machine Learning series). s.l.:s.n.

Jamie, D., 2012. How online reviews are crucial to a restaurant's takings. [Online]   
Available at: https://www.theguardian.com/lifeandstyle/2012/sep/02/ratings-boost-restaurants

Jeong, E. &. J., n.d. Restaurant experiences triggering positive electronic word-ofmouth (eWOM) motivations. International Journal of Hospitality Management.

Jingbo Zhu, H. W. M. Z. B. K. T. M., 2011. Aspect-Based Opinion Polling from Customer Reviews. IEEE Transactions on Affective Computing .

Keith Norambuena, B. |. L. E. F. |. V. C. M., 2019. Intelligent data analysis. [Online]   
Available at: https://content.iospress.com/articles/intelligent-data-analysis/ida173807

Lawson, R., 2015. Web Scraping by Python.

Lexalytics, 2020. https://www.lexalytics.com/technology/sentiment-analysis. [Online]   
Available at: https://www.lexalytics.com/technology/sentiment-analysis

Lexicon, O., n.d. Positive and Negative Words. [Online]   
Available at: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Luca, M. &. Z. G., 2016. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. [Online]   
Available at: https://doi.org/10.1287/mnsc.2015.2304

Lucas, M. B. D. a. R. E., 2013. The OxFord Handboook of Quantitative Methods in Psychology. s.l.:s.n.

Lu, X. B. S. H. L. F. Y., 2013. Promotional Marketing or Word-of-Mouth? Evidence from Online Restaurant Reviews. September.

M.N. Murty, ‎. R., 2016. Support Vector Machines and Perceptrons. In: s.l.:s.n.

Min Zhao, T. Z. & Chai, J., 2015. Based on SO-PMI Algorithm to Discriminate Sentimental Words. 2015 8th International Symposium on Computational Intelligence and Design (ISCID).

Mitchel, M., 2019. Artificial Intelligence: A Guide for Thinking Humans. In: s.l.:s.n.

Nicholas Cummins, S. A., 2018. MULTIMODAL BAG-OF-WORDS FOR CROSS DOMAINS SENTIMENT ANALYSIS. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Prabhu, T. N., 2019. EDA. [Online]   
Available at: https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce

python, 2020. scikit learn. [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

Python, n.d. URL REQUEST. [Online]   
Available at: https://docs.python.org/3/library/urllib.request.html

Rachmawan Adi Laksono, K. R. S. R. S. C. S. W., 2019. Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes. 12th International Conference on Information & Communication Technology and System (lCTS) 2019.

Relations, Y. I., 2020. https://www.yelp-ir.com/overview/default.aspx.

Rüdiger Wirth, J. H., 2004. CRISP-DM: Towards a Standard Process Model for Data Mining.

Sonawane, S. L. & Kulkarni, P. V., 2017. Extracting Sentiments from Reviews: A LexiconBased Approach. 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Issue 2017.

Stephanie, 2016. https://www.statisticshowto.com/research-methods-qualitative-research-and-quantitative-research/.

Stutton, R. S., 2018. Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning series). s.l.:s.n.

Sunil B. Mane, Y. S. S. K. V. S., n.d. Real Time Sentiment Analysis of Twitter Data.

Toth, A., 2010-2020. Scraping hub. [Online]   
Available at: https://www.scrapinghub.com/what-is-web-scraping/

Udemy, 2020. Udemy. [Online]   
Available at: https://blog.udemy.com/qualitative-vs-quantitative/?utm\_source=adwords&utm\_medium=udemyads&utm\_campaign=DSA\_Catchall\_la.EN\_cc.ROW&utm\_content=deal4584&utm\_term=\_.\_ag\_88010211481\_.\_ad\_437497337004\_.\_kw\_\_.\_de\_c\_.\_dm\_\_.\_pl\_\_.\_ti\_dsa-40209827789\_.\_li\_1007850\_

Van Dynk, B. E., 2007. Naive Bayesian classifiers for multinomial features: a theoretical analysis.

WebHarvy, 2016. https://www.webharvy.com/articles/what-is-web-scraping.html.

Wikipedia, n.d. https://en.wikipedia.org/wiki/Yelp#Features\_for\_businesses. [Online]   
Available at: https://en.wikipedia.org/wiki/Yelp#Features\_for\_businesses

Yang, F. X., 2017. Effects of restaurant satisfaction and knowledge sharing motivation on eWOM intentions: the moderating role of technology acceptance factors.. s.l.:s.n.

Yelp, 2004. Yelp.com. [Online]   
Available at: https://www.yelp-press.com/news/default.aspx  
[Accessed 2020].

Yelp, 2020. Yelp.com Dataset. [Online]   
Available at: https://www.yelp.com/dataset

Y, H. J. a. K., 2008. Journal of Korean Institute of Information Scientists and Engineers (KIISE): Computing Practices and Letters, 14, 336–340..