Sentiment Analysis of Restaurant reviews

By:

Bony Roy
CWID: 898161054

Girish Kumar Ramachandra
CWID: 888253630

CPSC 531 – 01 (13548)
Advanced Database Management System
Spring, 2019

Professor: Dr. Chun-I P. Chen

Department of Computer Science
California State University, Fullerton
May 7, 2019

Table of Contents

List of Figures

List of Tables

Abstract

Sentiment analysis is the process of interpreting an opinion from spoken or written language. In other words, it is identifying the emotional tone behind a sequence of texts or words. Sentiment analysis has allowed multinational companies to automate some of their important processes and get key insights to improve business. Nowadays most of the food joints or top restaurants want to know how customers think about the service quality and what steps they can take to improve their profit. Also, customers now dig into online reviews to decide from other's experience of service, food quality, discounts, and ambiance. Sentiment analysis of restaurant data using various algorithms can ease restaurant owners problem to a great extent. The process started by data collection from Yelp website and then various steps performed like data preprocessing, feature selection using TF-IDF, Training and test division. Then different algorithms such as Support Vector Machine(SVM), Naïve Bayes and multilayer perceptron(Artificial Neural network) applied on this dataset prior to final system implementation. The final system is built upon Support Vector Machine(SVM) which came out to be the best among three algorithms. SVM proved to be a very robust algorithm with high accuracy. The dataset divided into test and training set. Model is first trained using the training dataset and then it predicts the sentiment from the test dataset. Model validation is also an important part of sentiment analysis. Various model validation methods like holdout, Confusion Matrix and Classification Report applied. The system seems to predict positive and negative emotion reasonably well based on the dataset available. The future scope includes implementing a hybrid approach, training data on a larger dataset and implementing sarcasm detection.

Keywords: Sentiment Analysis, Sentiment Classification, Machine Learning, SVM, Neural network, Naïve Bayes.

Chapter 1: Introduction

1.1 Background

Availability of raw data was never an issue, how to make use of the raw data was always challenging. Today internet is in the zenith and "while social media plays a vital role in everyone's life in the form of microblogging sites, Twitter, Yelp, Facebook and other public media, is that the raw data is readily available for manipulation and analysis" (Kaur, Mangat & Nidhi, 2017). Sentiment analysis is a method of extracting opinionated, "selective data from the pool of raw data, where it allows us to know the views, attitudes, and emotions in public's tone" (Kaur et al.,2017). User's views and likes are collected and manipulated accordingly. This information further plays a major role in its analysis. Checking the attitude of the writer and his views about it is the task of sentiment analysis.

1.2 Problem Statements

Restaurant goers most of the time look upon online reviews before reserving a seat. They learn from others experience about restaurants ambiance, service, delivery time, worthiness, food quality. In yelp users post their rating and reviews on restaurants, businesses or services. So, Negative reviews from many users may have a large impact on potential customers who are making decisions. This is why business owners also nowadays look into these online forums to get an overall perspective about users opinion. Although there are many studies on yelp dataset, in the rapidly changing data world, hidden insights might be discovered from the dataset. Sentiment analysis categorizes the opinions into positive or negative or neutral view. This, in turn, will help users to make a decision and as well as the business owners' to improve their profit. Few points are important for restaurant owners' like if new customers are coming back, or existing customers increasing frequencies (Barbara Castiglia and Kevin Freibott, 2017). Hence the sentiment analysis system should be available to make business owners' as well as customers life easier. It will help in making a better decision from both points of views.

1.3 Project Goal and Benefits

- Analysis of reviews of restaurants based on customer's sentiments.
- Explore several supervised machine learning models that are used for sentiment analysis.
- Statistical analysis of the graphs and visualizations of Processed data related to top food joints in the USA.
- Prediction if a review is positive or negative based on the trained model.
- Based on the results the restaurants will always have the room for developments or modifications according to customers' expectations.
- Customers have a greater picture of the restaurants they dine in or they visit.

1.4 Relevance and Significance

The project associates and classifies the opinions expressed by the users on the website such as YELP in order to determine the restaurant goers or users attitude to be negative or positive. "The main purpose of Yelp is to provide a platform for customers to write a review along with providing a star-rating along with an open-ended comment and Yelp data is reliable, up-to-date and has a wide coverage of all kinds of businesses" (Boya et al.,2017). "Millions of people use yelp and empirical data demonstrated that Yelp restaurant reviews affected consumers' food choice decision-making" (Boya et al.,2017). Sentiment analysis provides splendid guidance to determine the performance of restaurants and to choose proper dining for customers which will improve customer satisfaction. It is important to comprehend the opinion based on the polarity.

1.5 Assumptions and Limitations

1.5.1 Assumptions

Due to the fact that the total dataset is 5GB which require more computing power. Data has been extracted for only a few restaurants or fast food joints. User can select a few restaurants (10 restaurants) to compare and get statistics of them. Based on processed data used to train the model, the computation increases exponentially. Complexity is influenced by multiple factors like Type of food, location, taste, service time, etc. Only textual reviews are considered and non-English reviews are not considered for being used in the project as proper supporting language dictionary might not be available.

1.5.2 Limitations

Classifying the sentiment from inadequate labeled data is a difficult task and even if efforts are made it's a costly process. Sometimes the reviews can be biased or posted from fake accounts.

Misspelling, short words, slangs or texts posted in Sarcastic tone may be found during the data preprocessing. Detecting these factors requires much more diverse dataset and relevant computing power. This project will overlook these limitations for the ease of the project.

1.6 Definition of Terms

Dataset – A collection of data is called dataset. As per the Cambridge dictionary, a dataset is a collection of separate sets of information that is treated as a single unit by a computer. A dataset might contain a stream of user reviews which includes user-name, location, ratings, reviews, or sentimental words over which the analysis is done.

Tokenization- Tokenization is, generally, an early step in the sentiment analysis process, a step which splits longer strings of text into smaller pieces, or **tokens**.

Feature: feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Pre-processing: An initial step in text and sentiment classification is pre-processing. A significant amount of techniques is applied to data in order to reduce the noise of text.

Chapter 2. Review of Literature

2.1 Sentiment Analysis

Sentiment analysis is a topic that fascinates many people mainly because it has many daily practical applications. Sentiment analysis is sometimes categorized as a subfield of "NLP (Natural language processing), is the computational handling of opinions, sentiments, feedback, and subjectivity of text"(Indiran, 2016). Various Sentiment analysis models parse online post or reviews and produce output in the form of polarities like –Neutral, negative or positive. The "overall process of sentiment analysis which starts from preprocessing of review dataset and continuous the sentiment classification or opinion mining through the various machine algorithms or dictionary-based techniques" (Abirami & Gayathri, 2017). Sentiment Classification techniques can be generally categorized into the lexicon-based approach and machine learning approach (Indiran, 2016). There is also another approach called a hybrid approach. Sentiment analysis can be done at various levels- we can divide it into three categories like Document level, sentence level and phrase level (Indiran, 2016).

2.1.1 Sentiment Analysis Process

Input dataset: First and foremost we need to gather data from a number of different sources for research.

Pre-processing: The purpose of Preprocessing is to process and represent the tweets in a cleaned and structured format. As most of the social network data are in the form of unstructured text, It helps to increase understanding of the emotion the text. It includes process like hashtag removal, URLs, repeated character removal, special symbol replacement, acronym and abbreviation expansions, and subject capitalization, etc. (Bhumika & Vimalkumar, 2016).

- Remove unwanted punctuation: Remove all unnecessary punctuation from the input text (Bhumika & Vimalkumar, 2016).
- Stop Word Removal: Pronouns (it, she/he), articles (the, a, an), prepositions (besides, in, near) are known as stop words. “They provide no or little information about sentiments and a list of stop words available on the internet, which can be used to remove them in the pre-processing step” (Kaur et al.,2017).
- Stemming: Stemming is basically removing suffixes and prefixes. For example, ‘working’, ‘worked’ can be stemmed to ‘work’. “It helps in classification but sometimes leads to decrease classification accuracy” (Kaur et al.,2017).
- Feature selection using TF-IDF: TF-IDF stands for Term Frequency and Inverse Document Frequency. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus(Wikipedia). TF(Term frequency) is basically the number of times term appears by the total number of terms in the document. IDF(Inverse Document Frequency) is the total number of documents(D) and number of documents with the term in it. It uses the formula as shown in figure 1.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 1: TF-IDF(Bag of Words & TF-ID. (n.d.))

The dataset from the preprocessing step is given as input to the different classification algorithms.

2.1.2 Supervised Machine Learning Approach

Supervised Machine Learning Approach (ML): Machine learning algorithms are combination techniques to automatically detect the hidden pattern in the dataset (Kaur et al.,2017). It makes use of undiscovered patterns to implement decision making under uncertainty or forecast future data (Kaur et al.,2017). The machine learning approach can be generally categorized into unsupervised and supervised learning methods. In the supervised method, the system is trained to make use of a large number of labeled training examples and in the Unsupervised method, the system is trained without labeled training documents (Indiran, 2016).

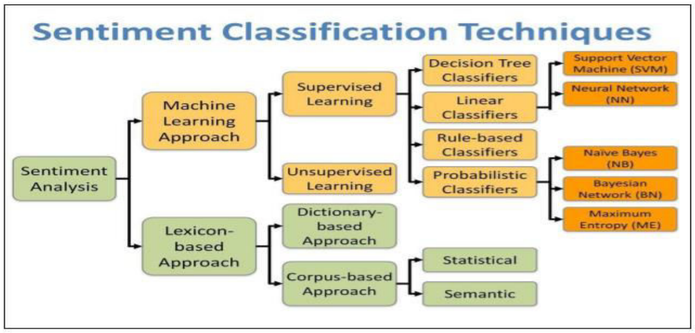


Figure 2: Sentiment Classification techniques (Bhumika & Vimalkumar, 2016)

Supervised Machine learning requires a training set to train the model and a test set to test the model. A training set consist of inputs of the system and expected outputs. A test set is previously not seen by the system and can be with new patterns. So, a test set is used to measure how accurate the model prediction is. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process.

2.1.2.1 Algorithm: Naive Bayes

Naïve Bayes is part of Probabilistic classifiers in Supervised Machine Learning approach. It is the most frequently used classifier due to its simplicity. This Classification theorem is named after "Thomas Bayes, who proposed the Bayes Theorem of determining the probability" (Parveen & Pandey, 2016). Bayesian classification or Naïve Bayes "provides useful learning algorithms and past knowledge and observed data can be combined with this" (Parveen & Pandey, 2016). It "assumes that the presence of a particular feature in a class is independent of the presence of any other feature" (Hlaing Moe et al., 2018).

It calculates the posterior probability P(C|X) from Prior probability of hypothesis C or P(C), evidence or predictor or Prior Probability of training data P(X) and the probability of X given C or P(X|C)" (Hlaing Moe et al., 2018). The "prediction result is the class with the highest posterior probability" (Hlaing Moe et al., 2018).

$$\boldsymbol{P(C|X)} = \frac{\boldsymbol{P(C)P(X|C)}}{\boldsymbol{P(X)}} \qquad \text{(Wikipedia)}$$

In simple English the equation can be written as:

$$\boldsymbol{Posterior} = \frac{\boldsymbol{Prior} \times \textbf{Likelihood}}{\boldsymbol{Evidence}} \qquad \text{(Wikipedia)}$$

For example, a fruit may be considered to be an apple if it is red, round, and about 3″ in diameter. A Naive Bayes classifier considers each of these "features" (red, round, 3" in diameter) to contribute independently to the probability that the fruit is an apple, regardless of any correlations between features. Features, however, aren't always independent which is often seen as a shortcoming of the Naive Bayes algorithm and this is why it's labeled "naive".

2.1.2.2 Algorithm: Support Vector Machine

Support Vector Machines (SVM) is a kind of linear classifiers in supervised learning. The "Support Vector Machine method is a classification approach which is based on the maximization of the margin" or distance between the separation hyperplane and instances (Amrani, Lazaar & Kadiri, 2018).

The main principle of the SVM training algorithm creates a model which categorizes new data into one or two classes(Kaur et al.,2017).Like: In figure 3, for example, we have three hyperplanes A, B, and C.

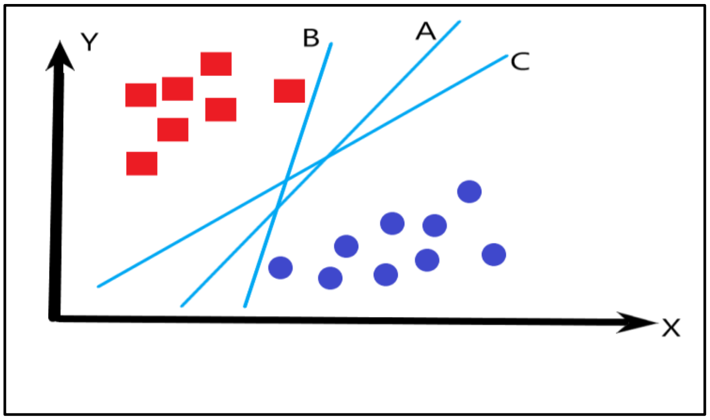


Figure 3: SVM Classification(Kaur et al.,2017)

We are going to classify Squares and Dots. The hyperplane C "provides the best separation between classes because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation" (Kaur et al.,2017). The hyperplane B provides the worst separation. So, to summarise "In this algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well" (Amrani et al., 2018). It is considered one of the best text classification methods(Amrani et al., 2018).

2.1.2.3 Artificial Neural Network(Multilayer Perceptron)

Artificial Neural network is a category of neural network algorithm based on artificial intelligence.

The neural network is similar to the neural structure of the brain and Neurons are the basic elements of this network(Kaur et al.,2017). “The inputs to the neurons are denoted by the vector overline Xi which is the word frequencies in the ith document. There is a set of weights A which are associated with each neuron used in order to compute a function of its' inputs f(). Multilayer neural networks are used for non-linear boundaries. These multiple layers are used to induce multiple piecewise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The output of the neurons in the earlier layers feed into the neurons in the later layers”(Indiran, 2016). “Deep learning on the term on neural network is the neural network with many of hidden layer in the system” (Ramadhani & Goo, 2017). First of all the review text is fed to the model for training which goes to different layers in an artificial neural network for classifying the output to be negative or positive. A multilayer perceptron (MLP) is a deep, artificial neural network. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP. The input text is sent

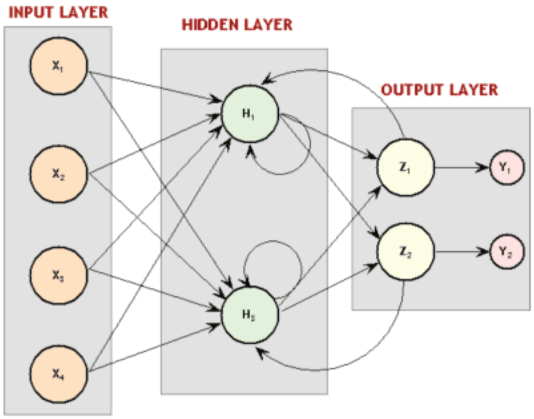


Figure 4: Neural Network

(Zharmagambetov & Pak, 2015)

to the input layer where every word is arranged in a look-up table in the form of vector or in other words Vectorization. Lastly, the output layer acts as a classifier model for classifying the output to be positive or negative.

2.2 Model Validation

This "process of deciding whether the numerical results quantifying hypothesized relationships between variables are acceptable as descriptions of the data, is known as validation"(Prashant Gupta, 2017). The training set is used by Machine learning to train their model. Therefore, the data used for training is important. If the model that has adequate training, the prediction is usually accurate. Model validation proves if a model is good or Bad. "Model validation provides a systemic way to measure accuracy and error rate which are commonly used to evaluate the performance of Machine Learning classification algorithms" (Omary & Mtenzi, 2010). The project uses Holdout method and confusion Matrix to validate the model.

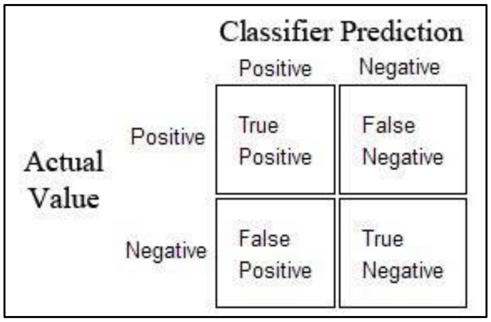
2.2.1 Holdout Method

Training and testing sets are separated into two non-overlapped groups. The basic method is "removing a part of the training data and using it to get predictions from the model trained on the rest of the data" (Prashant Gupta, 2017). The error estimation then tells how our model is doing on unseen data or the validation set. This is a "simple kind of cross-validation technique, also known as the holdout method" (Prashant Gupta, 2017). Data can be divided into 80:20,70:30, 75:25 etc. But "according to experts, 80:20 ratio, where 80% is for a training set and 20% is for testing, is ideal as each set is adequate for training and testing partitioning" (Cimentada, 2017).

So, in this project, we are following the suggestion of separating 80% of the dataset into a training set and rest 20% of the dataset into a testing set.

2.2.2 Confusion Matrix

A confusion matrix is used with classification models. The model predicts one response value for each observation in the testing set, and each predicted response value is compared to the actual response value for that observation present in the test set. So, confusion matrix assesses the accuracy of the predictive model built. Accuracy is "the percentage of correct prediction divided by the total number of predictions" (Kotsiantis, 2007). The numbers of correct and incorrect predictions in a model that classification technique is applied can be used to measure accuracy. Accuracy and error rate can be predicted by a confusion matrix. Figure 5 shows the confusion matrix where the rows represent the predicted class and columns represent actual values; True Positive means the positive examples are predicted correctly, False Positive means the positive examples are predicted incorrectly, False Negative means the negative examples are predicted incorrectly, and True Negative means the negative examples are predicted correctly.



		Classifier Prediction	
		Positive	Negative
Actual Value	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 5: Confusion Matrix (Banda et al.,2013)

Also, below 4 parameters are generated from this.

Accuracy: (TP+TN)/(TP+FP+TN+FN) (Wikipedia)

Precision measure accuracy of a class, when predicts "Yes" how often it is correct. Equal to TP/(TP+FP) (Wikipedia)

Recall: When it is actually "yes" how often it predicts "Yes". Equal to TP/(TP+FN). (Wikipedia)

F-measure: $2. \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$ (Wikipedia)

TP=True Positive **TN=**True Negative **FP=**False Positive **FN=**False Negative

Chapter 3: Methodology

3.1 Project methods, tools and techniques

This project sentimentally analyses the top restaurant's reviews from Yelp. The project uses Natural language processing for sentiment analysis, flask framework for web development and various machine learning algorithms.

Python: "**Python** is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace"(Wikipedia). In our project, we are using Python 3.7 which is open source and comes with IDLE(Integrated development platform).

MongoDB: **"MongoDB** is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemata"(Wikipedia).

3.2 Design and Functional Requirements

- The system will be taking data for processing from Mongo Db
- The system will have a dropdown menu with the top 10 food joint/Restaurant names.
- The system shall let the user select the name from the dropdown menu and submit it for further analysis.

- The system shall display the Sentiment Analysis of the restaurants with respective graphs and statistics
- The system shall display the confusion matrix for each food joints.
- The system will have another text area where user can add random restaurant reviews and it will predict if it is a positive text or negative text.
- The system will connect HTML for frontend with Python Backend with the help of the Flask framework of Python.
- The input textarea and foods joints dropdown will work individually.

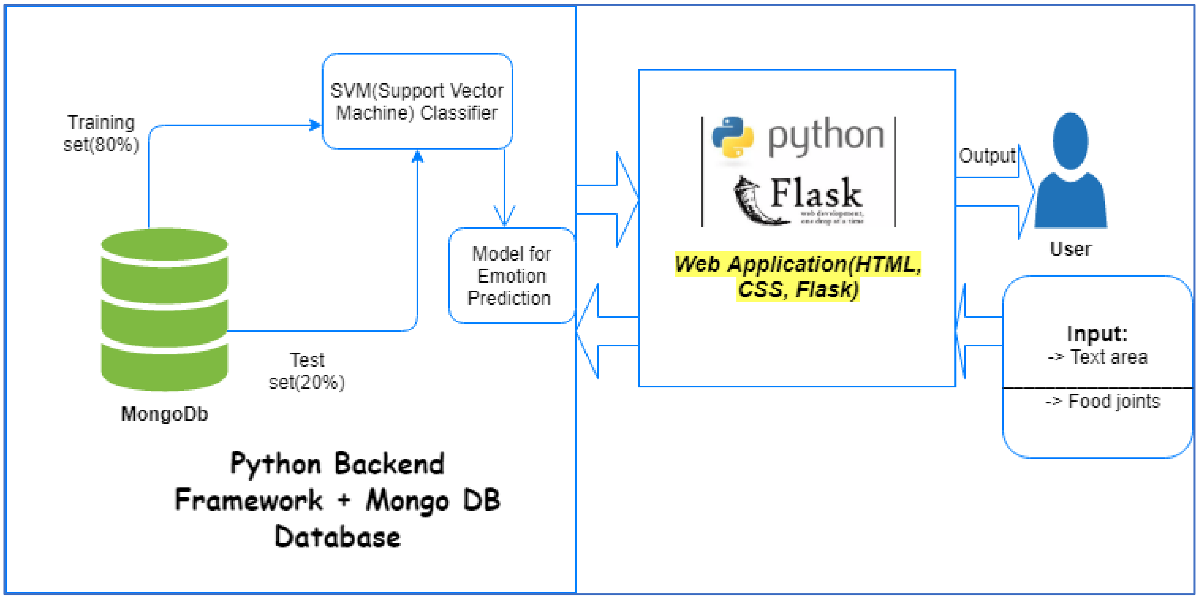

Training set(80%)
SVM(Support Vector Machine) Classifier
Model for Emotion Prediction
MongoDb
Test set(20%)
Python Backend Framework + Mongo DB Database
python
Flask
Web Application(HTML, CSS, Flask)
Output
User
Input:
-> Text area
-> Food joints

Figure 6: System Architecture Diagram

3.3 Data Requirement

There are two kinds of data in this project: data for training and data for real-time prediction of random restaurant reviews posted online.

3.3.1 Source of data collection, method of collection

Yelp Data set is downloaded from the Kaggle website directly in the form of a JSON file. Then the data is segregated to accumulate the reviews of top restaurants, which are then Sentimentally analyzed with respective. From website we have collected 2 datasets: “yelp_academic_dataset_business” and “yelp_academic_dataset_review”. The first set “yelp_academic_dataset_business” contains overall information(location, overall rating, ambiance etc.) about a restaurant or business. The second set of data contains an actual review about restaurants and date time. As mentioned before due to the file size we separated data for only 10 Restaurants/Food joints 2016 onwards. Then we have combined the “yelp_academic_dataset_business” and “yelp_academic_dataset_review” dataset to form a set of data with fields relevant to our project. Which as a whole comprise of nearly 20000 rows. The whole dataset is loaded into Mongo DB for faster processing and scalability of the system.

3.3.2 Inputs

The total yelp dataset is of 5 GB in size which way beyond our system computing capacity. So, we decided to take reviews of top 10 restaurants/food joints only from the year 2016 to 2018. This dataset comprises of nearly 20000 rows.

3.3.3 Reports

Users can select out of 10 restaurants to get the result; result consists of the sentimentally analyzed statistics of the respective restaurant. And the other data for real-time review should be directly typed in the comment section provided for this feature.

3.4 Data analysis and model to be used

3.4.1 Data Cleaning and Preprocessing

Dataset can be unstructured. It means that the data may contain an unnecessary component that the project does not need to use. The second step will be dedicated to the cleaning and preparing data for feature selection in the next step. From review text column we have removed URLs, Special characters, emoticons, etc. Then Stopwords are mainly collected from the corpus module of NLTK library in tagged to the English language. Stopwords are basically Pronouns (it, she/he), articles (the, a, an), prepositions (besides, in, near), etc.

3.4.2 Features Selection

Reviews can comprise of a lot of different words. The feature selection helps us to narrow them down to some key features that will be used for models. This step includes the TF-IDF vectorizer method. Some words that often appear will be eliminated by means of adding additional words into stopwords collection of NLTK library. Therefore, the feature selection usually removes other unnecessary words from the document besides from stopwords in the previous step.

3.4.3 Model Training

At this point, the models are developed. Each model uses different techniques and algorithms. 3 models being used, Naive Bayes, Support Vector Machine, and Artificial neural network(Multilayer perceptron) in this project. The whole dataset is divided into training and test set by the train_test_split module of sklearn library with 80(Training):20(Test). The training set is used for the model to be trained about the polarity of text. Then, the models are fed test dataset to test the accuracy.

3.4.4 Model Validation

A most important point of machine learning is accuracy. The prediction accuracy of the model decides if it is good and not. Various validation methods are used such as holdout method, (separate dataset into training and testing sets) and combine with the confusion matrix that helps to determine true of false percentage in the whole dataset. Each iterations average accuracy is calculated and integrated with the confusion matrix which then represents the model accuracy.

Also, the classification_report is generated with shows us Precision, Recall, F1, support. All these factors combined help up determine model accuracy.

3.4.5 Result Visualization and Analysis

The model predicts if the entered review is positive or negative. In the final system, we are displaying 10 restaurants in dropdown user can select any of them. The result shows the restaurants ratings(1/2/3/4/5) in a graphical representation, then for each restaurant selected the dataset is divided into test and training set and a confusion matrix is generated to show model accuracy. Also, with the chart of top 10 features or words that most of the reviews have mentioned and combined with top locations where users have given negative reviews with a count of reviews and top locations based on the count of positive reviews. Top features are selected based on the highest TF-IDF values. Most important part of the model is a real-time emotion prediction system. Where users can feed random restaurant reviews selected from internet and model will predict the provided review is positive sentiment or Negative sentiment.

3.5 System Requirements

3.5.1 Hardware Requirements

- Processor: Intel® Core™ i5-8250U CPU processor at 1.60 GHz or 1.80 GHz or 2.4 GHz
- RAM: 8GB
- Disk Space: 100 GB SSD
- GPU is preferred

3.5.2 Software Requirements

- Programming Language: Python 3.7(Opensource)
- Program: Python IDE such as Basic IDLE, PyCharm or Jupyter Notebook.(Opensource)

- Operating System: Windows 10(64 bit)
- Mongo Db Campus community(Opensource for local server)

3.6 System Installation and Technologies

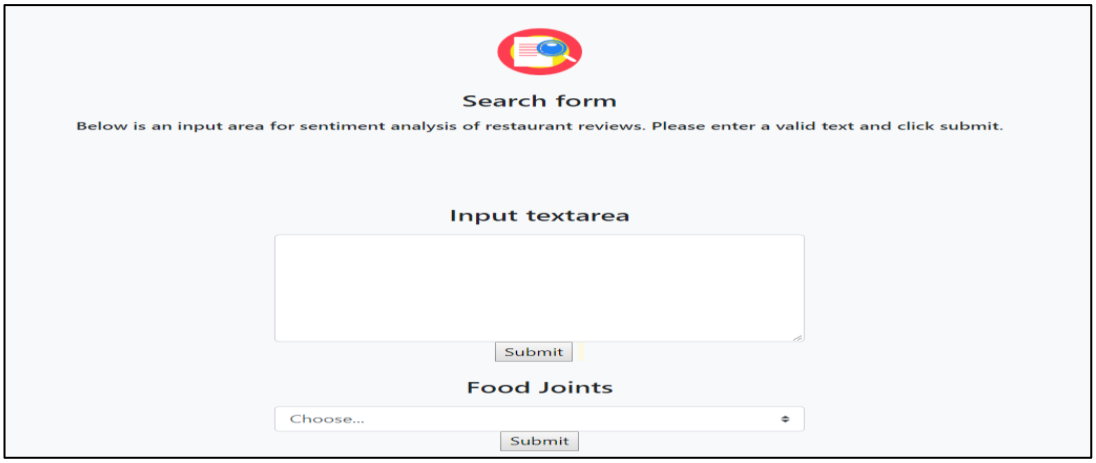
The system will be hosted in the local machine for the practice of this project. The project implemented mostly based on Python 3.7. All the libraries mentioned below are opensource. The followings describe the libraries from Python 3.7 is used during system installation and implementation:

- **JSON** - Support .json file read and writing
- **pandas** – Open source and very easy to use, data manipulation and analysis tool for Python Programming language. It is built on top of Numpy and key data structure is called data frame.
- **numpy** – Mathematical functions, multi-dimensional arrays, and matrices library
- **nltk** - Natural Language toolkit. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning.
- **Sklearn/Scikit-learn** – As per its website, tt is an Opensource, simple and efficient Machine learning library which contains various classification, regression, clustering algorithms. The main purpose of this library is to split data into test train and using machine learning algorithms on sets.
- **Matplotlib**- Python 2D plotting library
- **Seaborn**- Statistical data visualization
- **Flask**- a micro web framework for displaying the frontend.
- **Pymongo**- for connecting MongoDB server

Chapter 4: Results and Discussion

4.1 Results

The dataset that is used for training and test reviews are unstructured. They contain stopwords, punctuation, HTML elements, emoticons, special characters, uppercase, and lowercase combination and additional whitespaces. Therefore methods are implemented to remove this kind of noise from the dataset. Plus, this project applies the TF-IDF approach to select the features that are useful for training the models. These steps are prior to training the data with machine learning algorithms. Three machine learning models Multinomial NAÏVE BAYES, SVM and ANN(MLP classifer) are applied to the system for comparison. Then, the actual system is implemented with the best model, which is the Support Vector Machine(SVM). Confusion matrix along with Hold out method is used to measure the accuracy. The below page is the user interface of our system. Where the user can select a restaurant name from the dropdown and it will show relevant metrics of the restaurant. The input page also contains a text area where the user can enter a random restaurant review from internet and system will predict if the text is positive or negative.



Search form

Below is an input area for sentiment analysis of restaurant reviews. Please enter a valid text and click submit.

Figure 7: Sentiment Analysis system

Now, Suppose we have selected “Chipotle Mexican Grill” from the dropdown menu. The system will show different metrics related to those restaurants as shown in Figure 8,9,10,11.

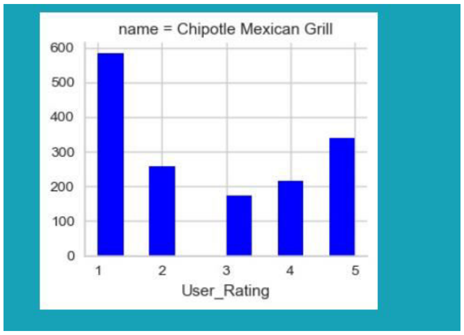


Figure 8: Comparison of the count of ratings of 5 categories.

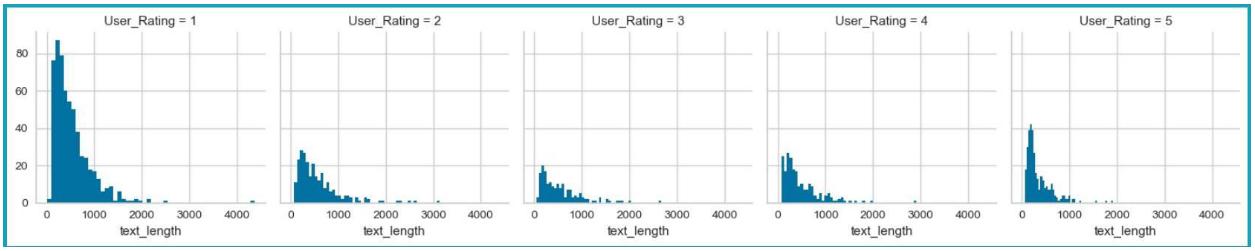


Figure 9: Average text lengths for each rating.

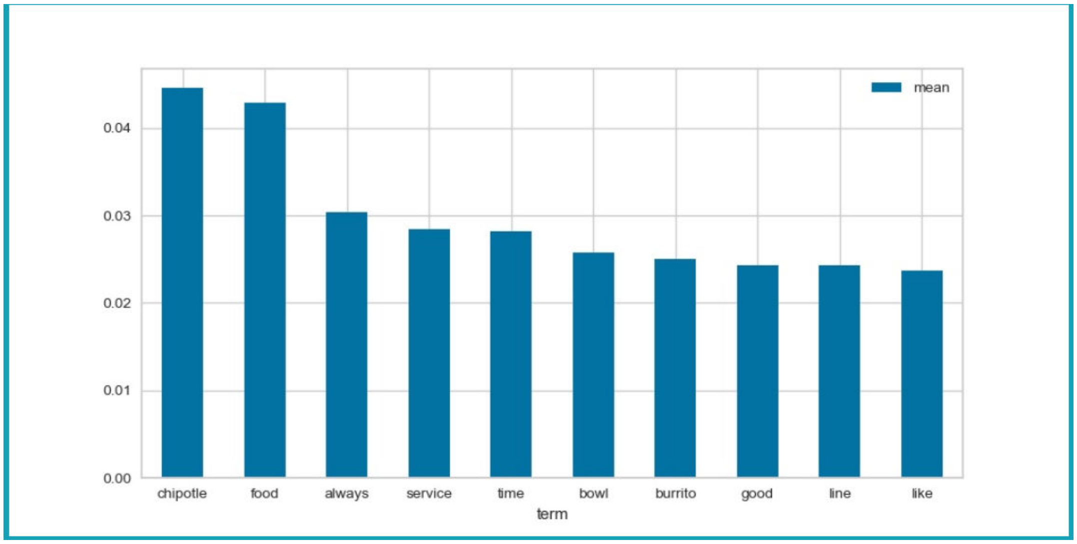


Figure 10: Feature matrix.

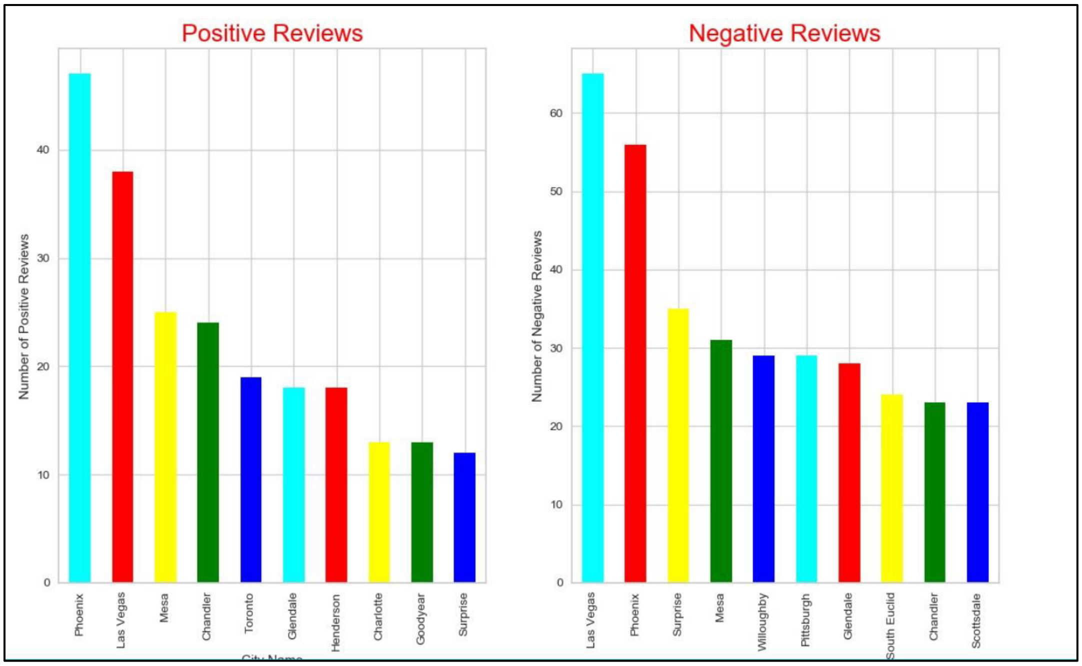
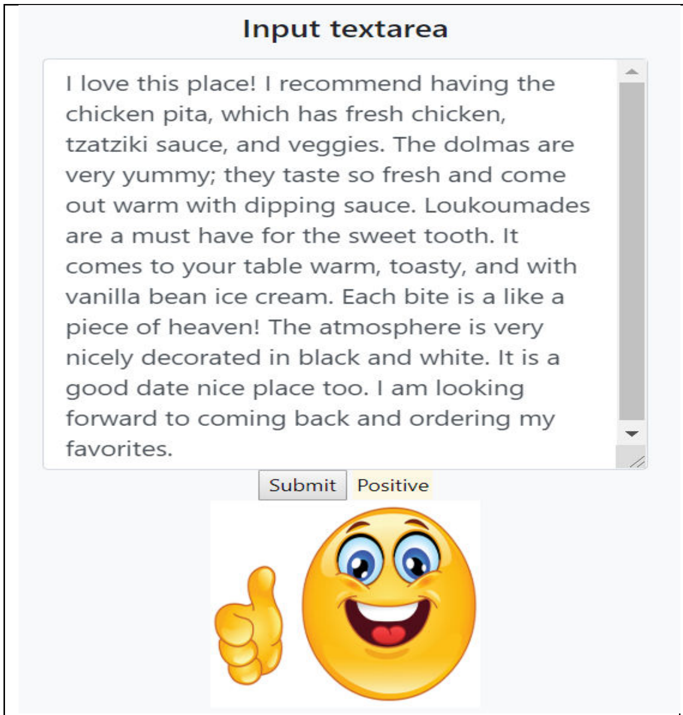


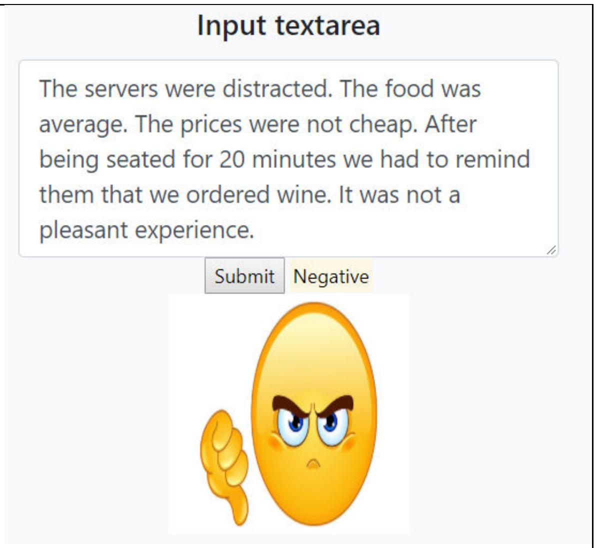
Figure 11: Locations with the most count of positive and negative reviews.

Below is the real-time prediction of the emotion of a text as shown in Figure 12 and 13.



I love this place! I recommend having the chicken pita, which has fresh chicken, tzatziki sauce, and veggies. The dolmas are very yummy; they taste so fresh and come out warm with dipping sauce. Loukoumades are a must have for the sweet tooth. It comes to your table warm, toasty, and with vanilla bean ice cream. Each bite is a like a piece of heaven! The atmosphere is very nicely decorated in black and white. It is a good date nice place too. I am looking forward to coming back and ordering my favorites.

Figure 12: Positive review prediction sample.



The servers were distracted. The food was average. The prices were not cheap. After being seated for 20 minutes we had to remind them that we ordered wine. It was not a pleasant experience.

Figure 13: Negative review prediction sample.

4.2 Discussion

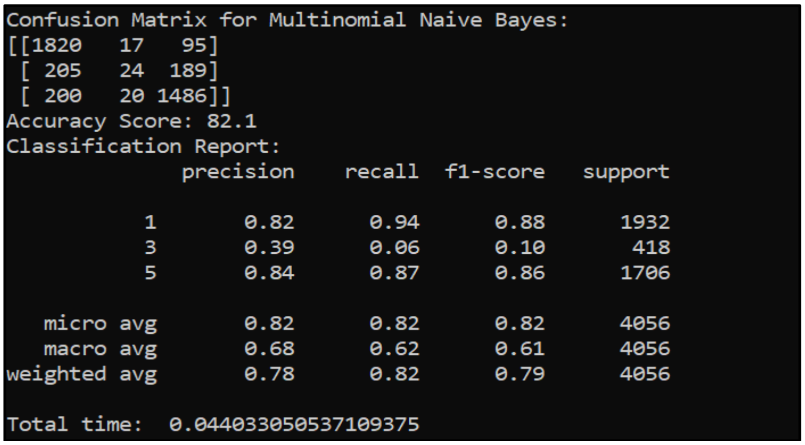
We collected all the datasets of these restaurants which nearly comprise of 20000 reviews. Divided it into test and training set and analyzed the result. For starter, we categorized rating 2 as negative review(or equivalent to rating 1) and rating 4 as positive review(or equivalent to rating 5). Now a review is positive if rated 5, neutral if rated 3 or Negative if rated 1. For the first analysis, we decided to choose all 3 categories(1/3/5 rating).

But the result did not meet our expectation. Below “table 1” shows the overall performance of 3 algorithms.

<table>
<tr>
<td>

Multinomial Naïve Bayes

</td>
<td>

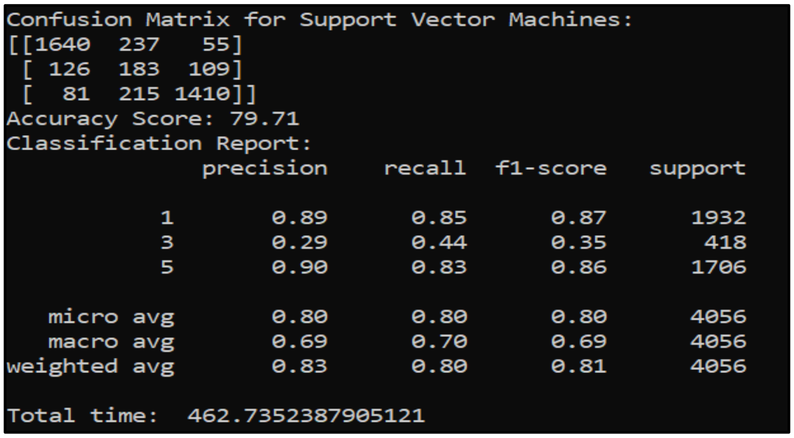


```
Confusion Matrix for Multinomial Naive Bayes:
[[1820   17   95]
 [ 205   24  189]
 [ 200   20 1486]]
Accuracy Score: 82.1
Classification Report:
              precision    recall  f1-score   support

           1       0.82      0.94      0.88      1932
           3       0.39      0.06      0.10       418
           5       0.84      0.87      0.86      1706

   micro avg       0.82      0.82      0.82      4056
   macro avg       0.68      0.62      0.61      4056
weighted avg       0.78      0.82      0.79      4056

Total time:  0.044033050537109375
```

</td>
</tr>
<tr>
<td>

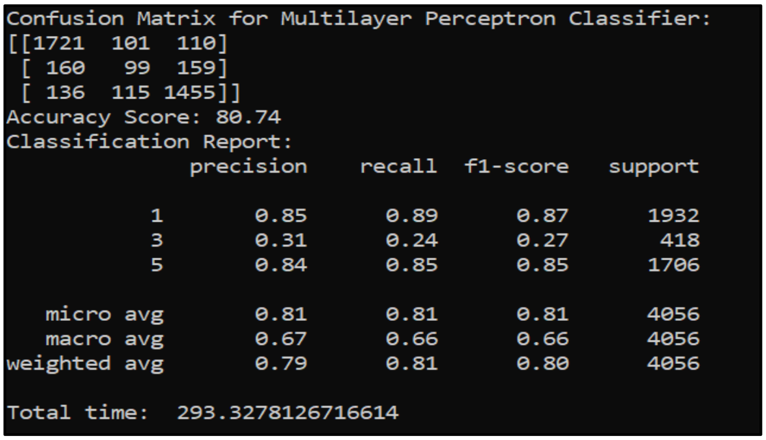
SVM(Support Vector Machine)

</td>
<td>



```
Confusion Matrix for Support Vector Machines:
[[1640  237   55]
 [ 126  183  109]
 [  81  215 1410]]
Accuracy Score: 79.71
Classification Report:
              precision    recall  f1-score   support

           1       0.89      0.85      0.87      1932
           3       0.29      0.44      0.35       418
           5       0.90      0.83      0.86      1706

   micro avg       0.80      0.80      0.80      4056
   macro avg       0.69      0.70      0.69      4056
weighted avg       0.83      0.80      0.81      4056

Total time:  462.7352387905121
```

</td>
</tr>
</table>

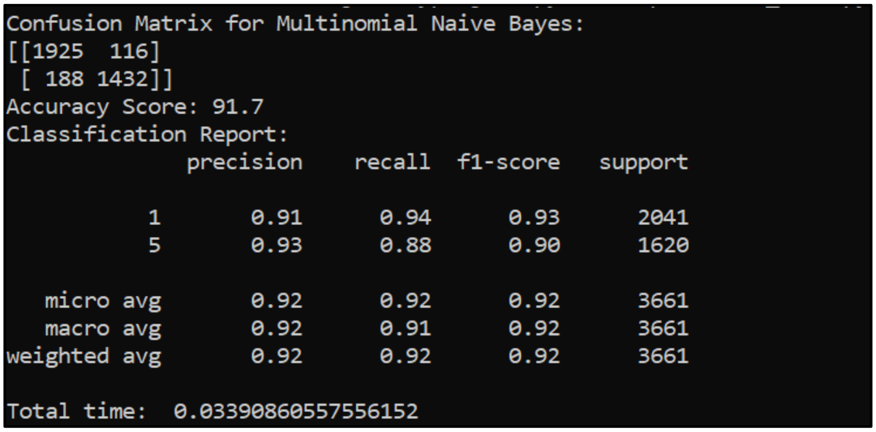
Multilayer **Perceptron(ANN)**	

```
Confusion Matrix for Multilayer Perceptron Classifier:
[[1721  101  110]
 [ 160   99  159]
 [ 136  115 1455]]
Accuracy Score: 80.74
Classification Report:
              precision    recall  f1-score   support

           1       0.85      0.89      0.87      1932
           3       0.31      0.24      0.27       418
           5       0.84      0.85      0.85      1706

   micro avg       0.81      0.81      0.81      4056
   macro avg       0.67      0.66      0.66      4056
weighted avg       0.79      0.81      0.80      4056

Total time:  293.3278126716614
```

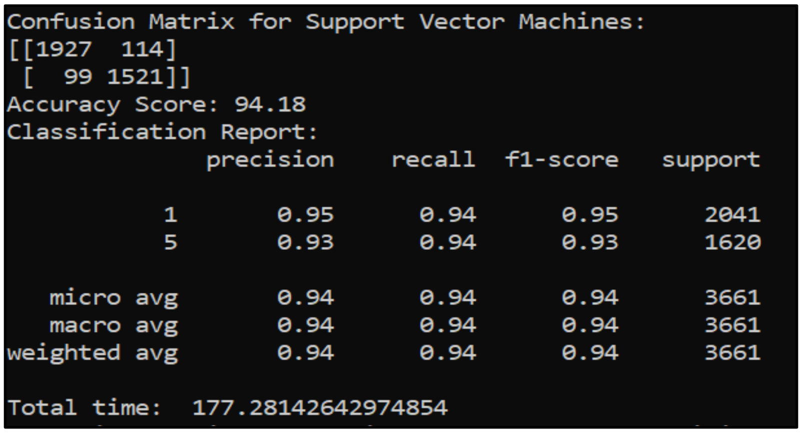
Table 1: comparison of 3 algorithms based on rating 1, 3 and 5.

Surprisingly, Naïve Bayes outperformed SVM and Multilayer both in terms of accuracy and total timing. But If we look closely at the result of prediction all the models predict the neutral(rating 3) with poor accuracy thus by reducing the accuracy of the model.

So, we decided to proceed with only 2 types of rating categories rating 5(Positive) and Rating 1(Negative) emotion. we also categorized rating 2 as equivalent to rating 1(Negative) and rating 4 as equivalent to rating 5(Positive). This time all the algorithms exceed our expectation and performed exceptionally well. **Table 2** is the main comparison criteria for this project.

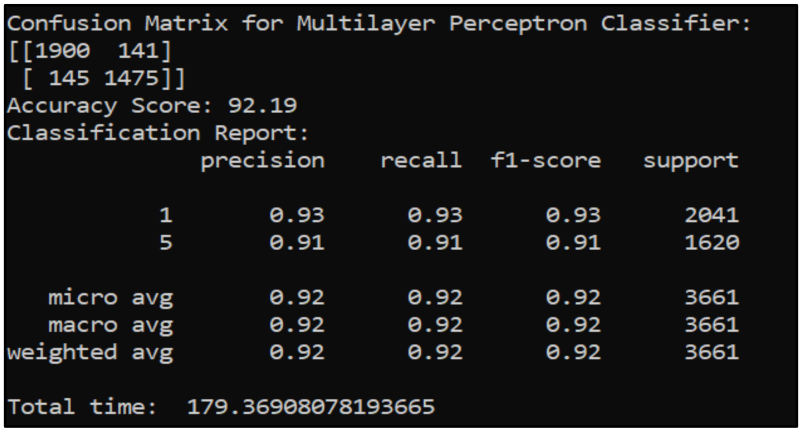
Multinomial **Naïve Bayes**	

```
Confusion Matrix for Multinomial Naive Bayes:
[[1925  116]
 [ 188 1432]]
Accuracy Score: 91.7
Classification Report:
              precision    recall  f1-score   support

           1       0.91      0.94      0.93      2041
           5       0.93      0.88      0.90      1620

   micro avg       0.92      0.92      0.92      3661
   macro avg       0.92      0.91      0.92      3661
weighted avg       0.92      0.92      0.92      3661

Total time:  0.03390860557556152
```

<table>
<tr>
<td><b>SVM(Support Vector Machine)</b></td>
<td>
<pre>
Confusion Matrix for Support Vector Machines:
[[1927 114]
 [99 1521]]
Accuracy Score: 94.18
Classification Report:
              precision   recall f1-score   support

           1   0.95   0.94   0.95   2041
           5   0.93   0.94   0.93   1620

   micro avg   0.94   0.94   0.94   3661
   macro avg   0.94   0.94   0.94   3661
weighted avg   0.94   0.94   0.94   3661

Total time: 177.28142642974854
</pre></td>
</tr>
<tr>
<td><b>Multilayer Perceptron(ANN)</b></td>
<td>
<pre>
Confusion Matrix for Multilayer Perceptron Classifier:
[[1900 141]
 [145 1475]]
Accuracy Score: 92.19
Classification Report:
              precision   recall f1-score   support

           1   0.93   0.93   0.93   2041
           5   0.91   0.91   0.91   1620

   micro avg   0.92   0.92   0.92   3661
   macro avg   0.92   0.92   0.92   3661
weighted avg   0.92   0.92   0.92   3661

Total time: 179.36908078193665
</pre></td>
</tr>
</table>

Table 2: comparison of 3 algorithms based on rating 1 and 5.

Naive Bayes does not perform very well comparing to Support Vector Machine in terms of accuracy and time. Therefore, the Support Vector Machine wins over Naive Bayes. On the other hand, Multilayer Perceptron(Artificial Neural Network) performs poorer than Support Vector Machine but better than Naïve Bayes. SVM outperforms Multilayer Perceptron(Artificial Neural Network) in terms of both time complexity and Accuracy. The ANN propagates through multiple hidden layers time complexity increases. This might be a disadvantage of ANN. Due to this reason, the project is decided to implement SVM rather than ANN.

The accuracy of Naïve Bayes is 91.7%, SVM is 94.18 % and MLP(Multilayer Perceptron- ANN based) is 92.19. In terms of time taken Naïve Bayes is much faster than both

SVM and Multilayer Perceptron(ANN). But accuracy is lowest among 3. SVM takes 177 seconds and Multilayer Perceptron(ANN) took 179 seconds. Although there is not much difference. But for datasets with higher amount of rows ANN might take longer time. Artificial Neural Networks and Deep Learning require more data for training. One main difference between deep learning/neural network and machine learning is the ability to extract features. With human intervention by feature extraction(by TF-IDF) Machine learning is able to predict the class. But ANN works similar to the human brain learns prediction and also feature extraction.

So, we can say based on evidence that SVM is the winner for our dataset. The final model which predicts the emotion(positive or negative) of a random restaurant review from the internet is based on this model.

Chapter 5: Conclusions, Implications, Recommendations

5.1 Conclusions

Sentiment analysis of yelp dataset applies supervised machine learning based approaches to predict emotion of a text. Several data preprocessing techniques applied to clean the data. Stopwords removed to increase the accuracy of the model. TF-IDF approach is mainly used for features selection. Machine Learning (Support Vector Machine) is used for the model training, prediction and model validation.

Several machine learning algorithms such as Naive Bayes and Multilayer Perceptron(ANN)has been applied, but SVM is chosen to be the best candidate for this system. SVM has the best performance in every area. Although Naïve Bayes is the fastest among all. Naïve Bayes has the poorest performance comparing to SVM and Multilayer Perceptron(ANN).

Since SVM offers better time efficiency and accuracy than Multilayer perceptron(ANN), it is no surprise if SVM can beat Multilayer perceptron(ANN) in the domains that time is critical. Naïve Bayes and SVM are machine learning algorithms. Multilayer perceptron(ANN) is a deep learning model (which is a subset of machine learning that does not contain SVM or Naïve Bayes). Naïve Bayes applies the essential and age-old probability theorem, SVM applies optimal hyperplane of separation, and ANN applies the concept of neurons of the human brain to predict the outcomes.

The system seems to work decently as it predicts sentiment for random reviews pretty accurate. The system performs acceptably if only the purpose of the project is considered. Although it performs well but accuracy can be increased by feeding more data.

5.2 Lesson Learned

There are many benefits to this project. Someone who truly knows the meaning and sentiments of a system should label the dataset. The dataset of Yelp is nearly 5GB. This large dataset requires much more computing power. Our project system has limited RAM of 8 GB. Which does not suffice this large dataset. Due to this constraint, we only worked on a set of data. So, for future implementation, it will be great if we can include the whole dataset. Sentiment analysis can be combined with other metrics. More time and Large dataset are required for the accuracy and much correct prediction of emotion.

5.3 Implications

Nowadays Sentiment analysis is used everywhere, especially the fields or businesses that rely on reviews. like games reviews, restaurant reviews, hospital industry, books reviews, and movies reviews. Business value can be improved by these. As restaurants come up with different themes and menus every day. Writing reviews will help to a greater extent. Not only the user makes the decision easier for others to get a glimpse of the restaurants but also allows restaurant management to improve their service, that is another perk of the system. In terms of application development, Flask framework seems to easily integrate an HTML template with Python Backend. Understand the importance of various python libraries like SKlearn, Pandas, Numpy, Matplotlib, Seaborn. Utilizing MongoDB for Json type(Non-structured) data loading and processing. Jupyter Notebook can also be utilized to combine all of these and visually displaying the result.

5.4 Recommendations/Future Scope

- The sentiment analysis currently performed based on the historical dataset. In the future, the system can be implemented to utilize real-time data from APIs. That will in turn, help us to give a more accurate result.
- Currently unable to detect Neutral emotion from a Text. In Future scope, Neutral emotion detection can be done by incorporating much larger dataset.
- Researchers have pointed out we can use a hybrid classification methodology. In which we can combine classification methods like random forest and support vector machine to generate a new classification technique. The hybrid approach sometimes can improve accuracy than the simple classifier models. In future scope, a hybrid approach can be implemented to build more robust, fast and accurate classifiers.
- An increasing amount of Sarcasm can be seen in posts nowadays. But the important point is, extracting proper emotion from these sarcastic posts can be an overwhelming task for classifier algorithms. So, we could enhance the system in a way that it can detect sarcasm from text and classify the text in the correct context.

References

Barbara Castiglia and Kevin Freibott (2017). What Data Points Are Important For Restaurants. Retrieved from https://www.modernrestaurantmanagement.com/what-data-points-are-important-for-restaurants/.

Bag of Words & TF-ID. (n.d.). Retrieved from https://skymind.ai/wiki/bagofwords-tf-idf

Omary, Z., &amp;Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. Retrieved from https://www.researchgate.net/publication/228548543_Machine_Learning_Approach_to_Identifying_the_Dataset_Threshold_for_the_Performance_Estimators_in_Supervised_Learning

Indiran, Mohan (2016). Survey-Algorithms Used For Sentiment Analysis. Retrieved from https://www.researchgate.net/publication/317004065_Survey-Algorithms_Used_For_Sentiment_Analysis

A. M. Abirami and V. Gayathri(2017). A survey on sentiment analysis methods and approach. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7951748&isnumber=7951734

H. Kaur, V. Mangat and Nidhi(2017). A survey of sentiment analysis techniques. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8058315&isnumber=8058234

H. Parveen and S. Pandey (2016). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7912034&isnumber=7911953

M, Bhumika & B, Vimalkumar (2016). Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis. Retrieved from https://www.researchgate.net/publication/305361796_Sentiment_Analysis_using_Support_Vector_Machine_based_on_Feature_Selection_and_Semantic_Analysis

Hlaing Moe, Zun & San, Thida & Mie Khin, Mie & May Tin, Hlaing (2018). Comparison Of Naive Bayes And Support Vector Machine Classifiers On Document Classification. Retrieved from https://www.researchgate.net/publication/329654958_Comparison_Of_Naive_Bayes_And_Support_Vector_Machine_Classifiers_On_Document_Classification

Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri(2018). Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877050918301625

A. M. Ramadhani and H. S. Goo(2017). Twitter sentiment analysis using deep learning methods. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8068556 &isnumber=8068531

A. S. Zharmagambetov and A. A. Pak(2015), Sentiment analysis of a document using deep learning approach and decision trees. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp ?tp=&arnumber=7416902&isnumber=7416865

Prashant Gupta(2017). Cross-Validation in Machine Learning. Retrieved from https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Cimentada, J. (2017, Sep). Holdout and cross-validation. Retrieved from https://cimentadaj.github.io/blog/2017-09-06-holdout-and-crossvalidation/holdout-and-crossvalidation/

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Retrieved from https://datajobs.com/data-science-repo/

Banda, Juan & Angryk, Rafal & Martens, Petrus. (2013). Steps Toward a Large-Scale Solar Image Data Analysis to Differentiate Solar Phenomena. Retrieved from https://www.researchgate.net/publication/256418526_Steps_Toward_a_Large-Scale_Solar_Image_Data_Analysis_to_Differentiate_Solar_Phenomena

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the ACL. Retrieved from https://www.cs.cornell.edu/home/llee/papers/cutsent.pdf.

Yu, Boya & Zhou, Jiaxu & Zhang, Yi & Cao, Yunong. (2017). Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. Retrieved from https://www.researchgate.net/publication/320055668_Identifying_Restaurant_Features_via_Sentiment_Analysis_on_Yelp_Reviews