

CPSC 481-Artificial Intelligence (Assignment-5-Project Report)

Project: Sentiment Analysis of Restaurant Reviews

(GIRISH KUMAR RAMACHANDRA, 888253630)

❖ **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 owner's problem to a great extent. The process started by data collection from Yelp website and then various steps performed like data pre-processing, 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

❖ **Sentiment Analysis.**

Sentiment Analysis is a process where we can classify and identify several feelings, opinions and classify them based on raw opinionated data, in which we identify the attitude of a person towards a view or product and the basic idea is to mainly find the bias or view and classify them accordingly into positive view , negative view or not biased (Neutral). It is an application of natural language processing. also sometimes known as opinion mining or emotion extraction. Text mining is a very famous research field of Sentiment analysis. I shall be making use of Yelp Data sets in order to build a Sentimental Analysis project.

❖ **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.

❖ **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 Pre-processing 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.

Remove unwanted punctuation: Remove all unnecessary punctuation from the input text.

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".

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".

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. 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

The dataset from the pre-processing step is given as input to the different classification algorithms.

❖ System Requirements

✓ 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

✓ **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)

❖ **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, it 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

❖ **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”. Bayesian classification or Naïve Bayes “provides useful learning algorithms and past knowledge and observed data can be combined with this”. It “assumes that the presence of a particular feature in a class is independent of the presence of any other feature”.

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 $P(X|C)$. The “prediction result is the class with the highest posterior probability”.

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)}$$

In simple English the equation can be written as:

$$\textbf{Posterior} = \frac{\textbf{Prior} \times \textbf{Likelihood}}{\textbf{Evidence}}$$

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 labelled “naive”.

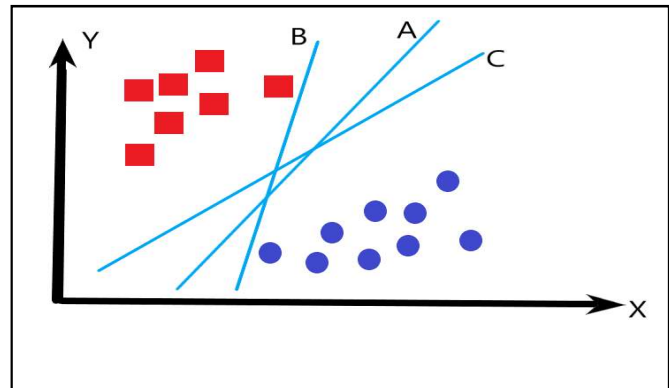
❖ Algorithm: Support Vector Machine

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.

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

Figure 3: SVM Classification

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”. 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”. It is considered one of the best text classification methods.

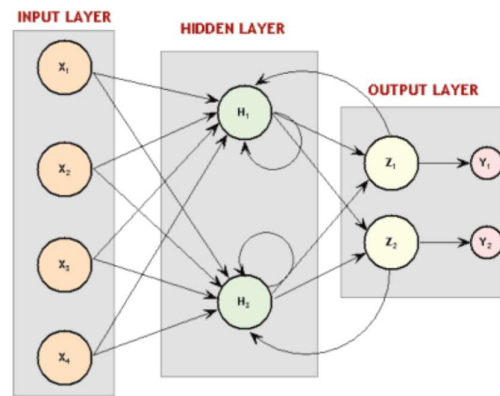
❖ Artificial Neural Network (Multilayer Perceptron)

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. “The inputs to the neurons are denoted by the vector $\overline{X_i}$ which is the word frequencies in the i^{th} 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". "Deep learning on the

Figure 4: Neural Network

term on neural network is the neural network with many of hidden layer in the system".



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 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.

❖ 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". 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”. The project uses Holdout method and confusion Matrix to validate the model.

✓ **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”. 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”. 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”.

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.

✓ **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”. 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

Also, below 4 parameters are generated from this.

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

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

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

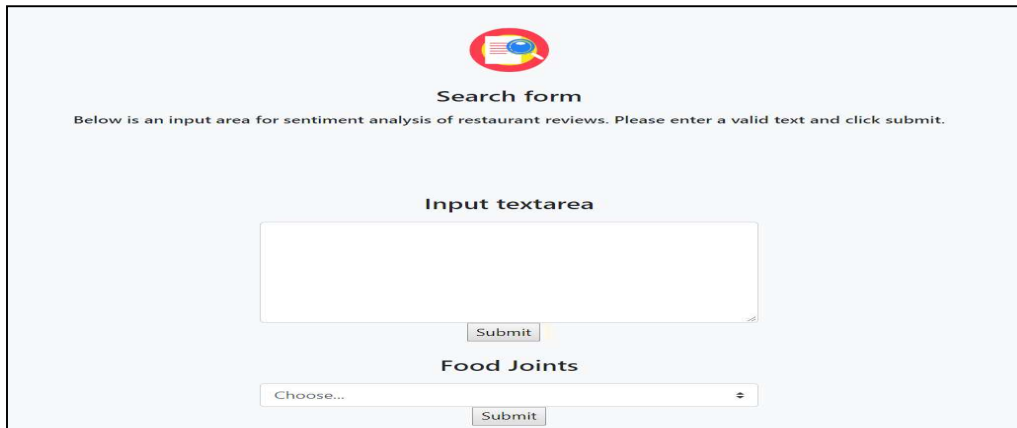
F-measure: $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Rec}}$

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

❖ Results

The dataset that is used for training and test reviews are unstructured. They contain stop words, 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 classifier) 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.



The screenshot shows a web interface for a sentiment analysis system. At the top, there is a logo with a magnifying glass over a document. Below the logo, the text "Search form" is displayed. Underneath, a message states: "Below is an input area for sentiment analysis of restaurant reviews. Please enter a valid text and click submit." In the center, there is a large text input area labeled "Input textarea". Below the input area is a "Submit" button. At the bottom, there is a dropdown menu labeled "Food Joints" with the text "Choose..." and a "Submit" button.

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.

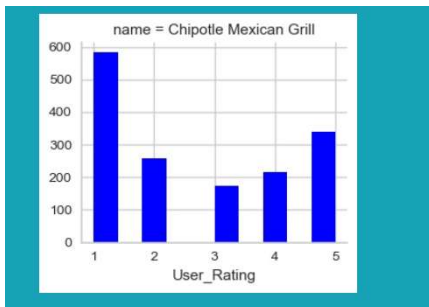


Figure 8: Comparison of the count of ratings of 5 categories.

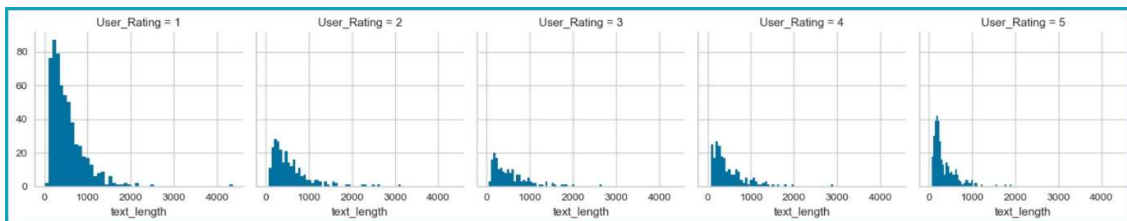


Figure 9: Average text lengths for each rating.

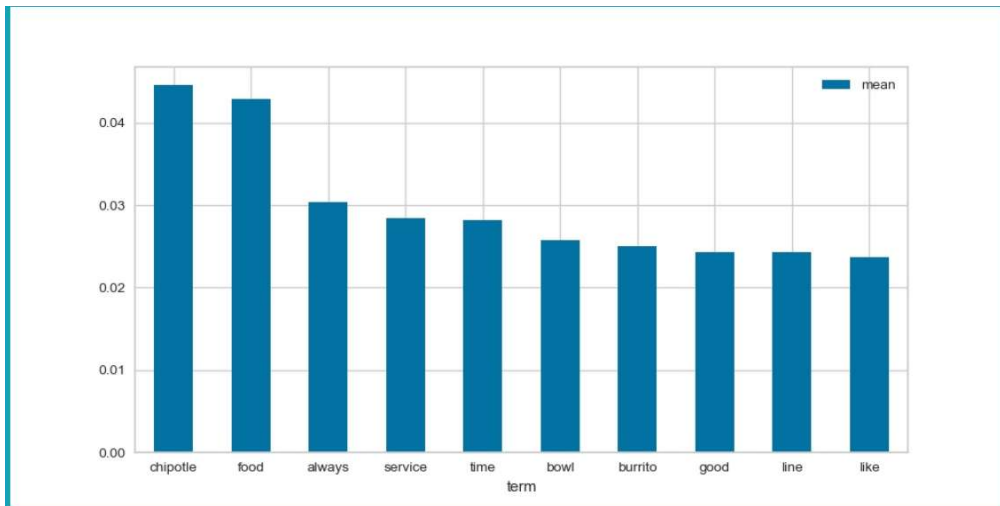


Figure 10: Feature matrix.

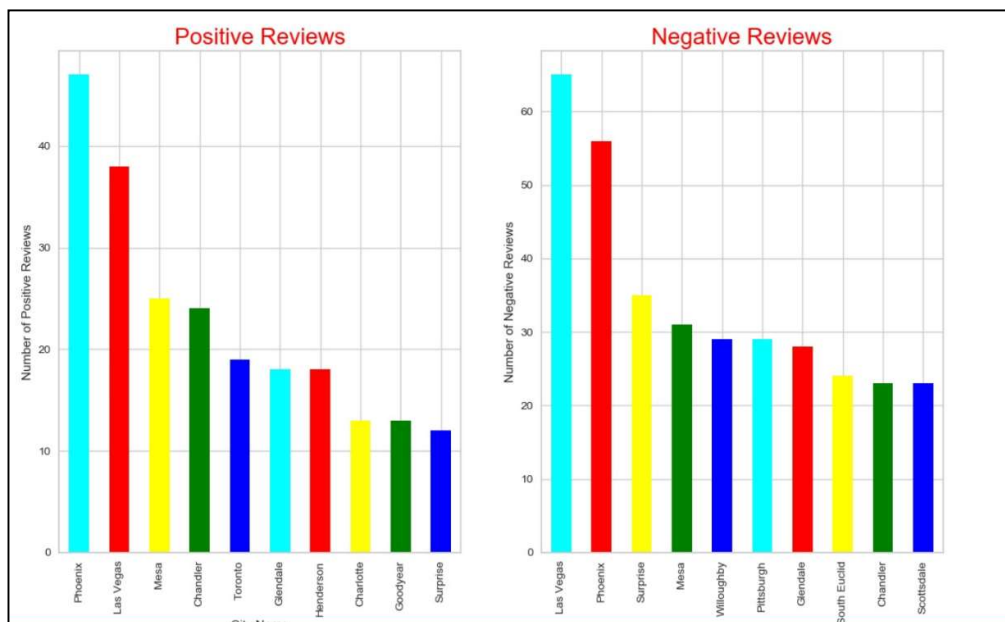
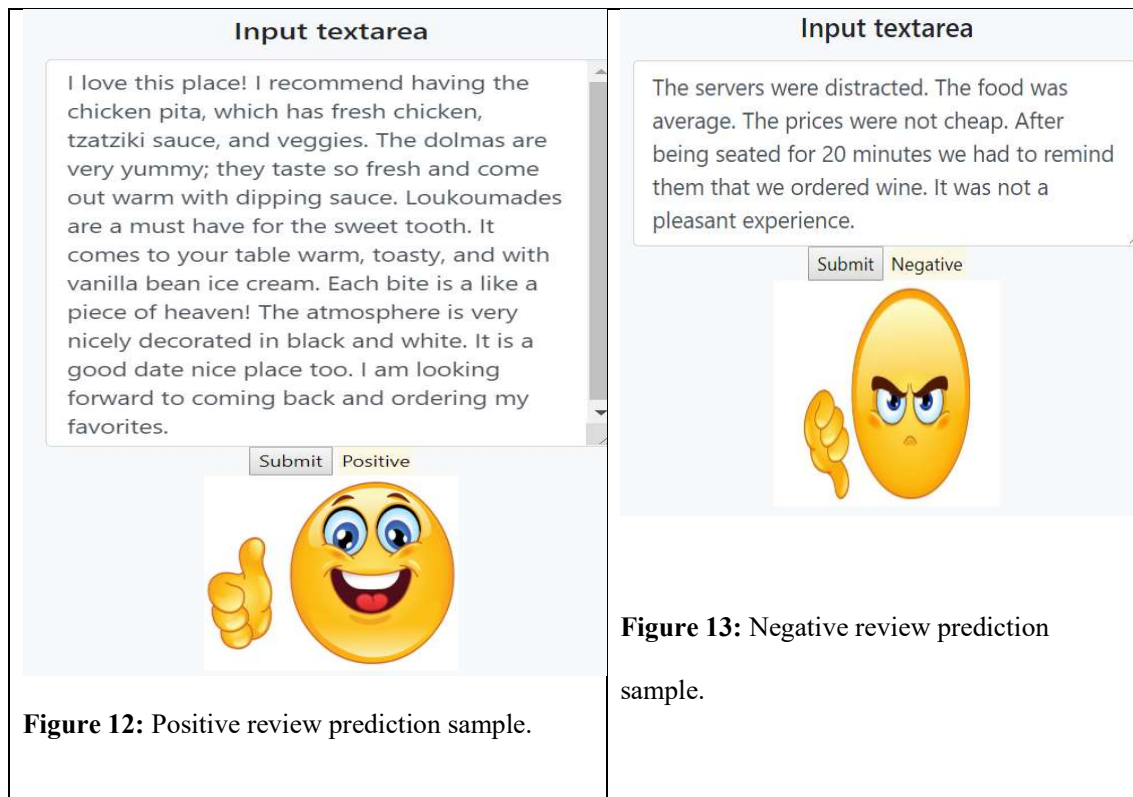


Figure 11: Locations with the most count of positive and negative reviews.

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



❖ Discussion

We collected all the datasets of these restaurants which nearly comprise of 20000 reviews. Divided it into test and training set and analysed 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.

Multinomial Naïve Bayes	<pre> 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: 0.0444033050537109375 </pre>
SVM(Support Vector Machine)	<pre> 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 </pre>
Multilayer Perceptron(ANN)	<pre> 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 </pre>

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	<pre> 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 </pre>
SVM(Support Vector Machine)	<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>
Multilayer Perceptron(ANN)	<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>

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.

❖ **Conclusions**

Sentiment analysis of yelp dataset applies supervised machine learning based approaches to predict emotion of a text. Several data pre-processing techniques applied to clean the data. Stop words 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 Naïve 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.

❖ **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.

❖ **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.