**A project report**

**on**

**SENTIMENT ANALYSIS BASED ON NATURAL**

**LANGUAGE PROCESSING**

Submitted in partial fulfillment for the award of the degree of

Bachelor of Engineering

in

Computer Science and Engineering

by

**RUISO MALUE (D/18/CS/108)**

**RINCHUI CHIPHANG (D/18/CS/109)**

*under the guidance of*

Mr. Ajit Kr. Singh Yadav

(Assistant Professor)



**DEPARMENT OF COMPUTER SCIENCE AND ENGINEERING**

**North Eastern Regional Institute of Science and Technology** (**Deemed to be University)**

**Nirjuli, Arunachal Pradesh-791109**

June 2022

**DECLARATION**

This is to certify that project entitled “**SENTIMENT ANALYSIS BASED ON NATURAL LANGUAGE PROCESSING**” which is submitted by us in partial fulfillment of the requirement for the award of degree in Computer Science & Engineering at North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh comprises only our own work and due acknowledgement has been made in the text to all other material used.

Date:

Name of the Student: Signature:

Ruiso Malue (D/18/CS/108)

Rinchui Chiphang (D/18/CS/109)



**CERTIFICATE OF APPROVAL**

Certified that the project report entitled “**SENTIMENT ANALYSIS BASED ON NATURAL LANGUAGE PROCESSING”** is a bonafide work carried out jointly by **Ruiso Malue (D/18/CS/108) and Rinchui Chiphang (D/18/CS/109).** The project report embodies the original work done by them towards partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** at **North Eastern Regional Institute of Science and Technology, Arunachal Pradesh**. It is understood by this approval that the undersigned do not endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project report only for the purpose for which it has been submitted.

Ajit Kr. Singh Yadav Ashwini Patra Kumar

( (Project Supervisor) (Project Co-ordinator)

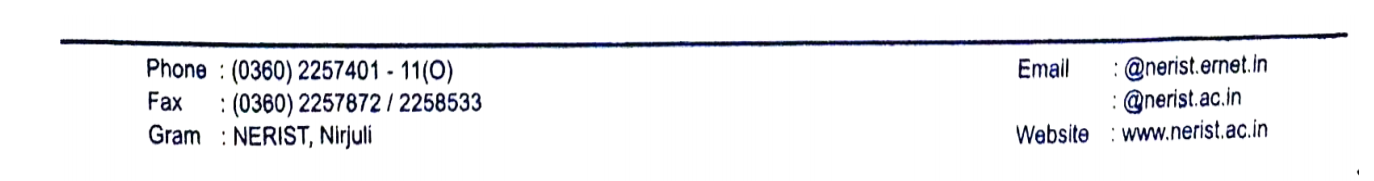
Asst. Professor Asst. Professor

Department of CSE Department of CSE

Dr. Moirangthem Marjit Singh

Associate Professor

Head of Department, CSE.



**ACKNOWLEDGEMENT**

A successful and satisfactory completion of any project is the outcome of invaluable and aggregate contribution of different personal pulls in radial directions, explicitly or implicitly.

Like every big thing, this project is not only our effort. A lot of others who supported us in various ways, while some in technical way, and some, in moral way.

We would like to express our sincere gratitude to our Project guide **Mr. Ajit Kr. Singh Yadav**, Assistant Professor, Dept. of Computer Science & Engineering, NERIST who has helped us a lot in providing all facilities needed and proper guidance and coordination in completing the project within specified time for the completion of this project.

The suggestions and criticism of these people significantly improved the project.

Name of the student: Signature:

Ruiso Malue (D/18/CS/108)

Rinchui Chiphang (D/18/CS/109)

**ABSTRACT**

A dataset of about thousand review given by the customer of the restaurant. Our aim is to classify the reviews with different kinds of sentiment into positive or negative review. Using five different algorithm a dataset is trained. The trained dataset is then used to predict the other test set. The five machine learning algorithm classifier are then compare Based on the accuracy score, precision and recall to get the best fitted model for the review dataset.

**TABLE OF CONTENT**

[**Chapter 1 1**](#_Toc106656656)

[**INTRODUCTION 1**](#_Toc106656657)

[1.1 Machine learning 1](#_Toc106656658)

[1.2 Types of machine learning 1](#_Toc106656659)

[1.2.1 Supervised machine learning 1](#_Toc106656660)

[1.2.2 Unsupervised Machine Learning 1](#_Toc106656661)

[1.2.3 Reinforcement Learning 2](#_Toc106656662)

[1.3 Application of Machine Learning 2](#_Toc106656663)

[**Chapter 2 3**](#_Toc106656664)

[**LITERATURE SURVEY 3**](#_Toc106656665)

[2.1 Natural Language Processing (NLP) 3](#_Toc106656666)

[2.2 NLP Tasks 3](#_Toc106656667)

[2.3 Sentiment Analysis 3](#_Toc106656668)

[2.4 Types of sentiment analysis 4](#_Toc106656669)

[2.5 Sentiment Analysis Applications 4](#_Toc106656670)

[2.6 Sentiment Analysis Challenges 4](#_Toc106656671)

[2.7. Some basic terminology 5](#_Toc106656672)

[**Chapter 3 7**](#_Toc106656674)

[**METHOLOGY 7**](#_Toc106656675)

[3.1 Logistic Regression 7](#_Toc106656676)

[3.2 Naïve Bayes Model 7](#_Toc106656678)

[3.3 Support Vector Machine 8](#_Toc106656679)

[3.4. Decision Tree 9](#_Toc106656680)

[3.5. Random Forest 10](#_Toc106656681)

[3.6 Software Description 11](#_Toc106656682)

[3.7 Software requirements 11](#_Toc106656683)

[**Chapter 4 12**](#_Toc106656684)

[**DESIGN 12**](#_Toc106656685)

[4.1 Design Flow 12](#_Toc106656686)

[4.1.1. Training data 13](#_Toc106656687)

[4.1.2. Features 13](#_Toc106656688)

[4.1.3. Model 13](#_Toc106656689)

[4.1.4. Input 13](#_Toc106656690)

[4.1.5. Label 13](#_Toc106656691)

[**Chapter 5 14**](#_Toc106656692)

[**IMPLEMENTATION 14**](#_Toc106656693)

[5.1. Implementation flowchart 14](#_Toc106656694)

[5.2. Review Dataset 15](#_Toc106656695)

[5.3. Data Preprocessing 16](#_Toc106656696)

[5.3.1. Punctuation 16](#_Toc106656697)

[5.3.2. Tokenization 16](#_Toc106656698)

[5.3.3. Stop words 16](#_Toc106656699)

[5.3.4 Stemming 16](#_Toc106656700)

[5.4. Transformation 17](#_Toc106656701)

[5.4.1 Creating Bags of words 17](#_Toc106656702)

[5.5. Splitting of dataset into Training and Test set 19](#_Toc106656703)

[5.6. Training of Dataset 20](#_Toc106656704)

[5.6.1.Logistic Regression 20](#_Toc106656705)

[5.6.2. Naïve bayes 20](#_Toc106656706)

[5.6.3. SVM 20](#_Toc106656707)

[5.6.4. Decision tree 21](#_Toc106656708)

5.6.5. Random forest………………………………………………………….............24

[**Chapter 6 22**](#_Toc106656709)

[**RESULT AND ANALYSIS 22**](#_Toc106656710)

[6.1. Results based on Confusion Matrix 22](#_Toc106656711)

[6.1.1. Naïve Bayes Model 22](#_Toc106656712)

[6.1.2. Logistic Regression 22](#_Toc106656713)

[6.1.3. Random forest 23](#_Toc106656714)

[6.1.4. Support Vector Machine 23](#_Toc106656715)

[6.1.5. Decision Tree 24](#_Toc106656716)

###### 6.2. Accuracy comparison between Models before and after Cross validation…………..25

[6.3. Precision and Recall comparison between the models 25](#_Toc106656718)

[6.4. Analysis 26](#_Toc106656719)

[**Chapter 7 27**](#_Toc106656720)

[**CONCLUSION AND FUTURE SCOPE 27**](#_Toc106656721)

[**8. REFERENCES 28**](#_Toc106656722)

**LIST OF FIGURES**

[Fig 3.1 Standard logistic function (𝑡)………. 7](#_Toc106657655)

[Fig 3.4. Support Vector Machine 8](#_Toc106657656)

[Fig 3.5. General structure of Decision Tree 9](#_Toc106657657)

[Fig 3.6 Random Forest Structure 10](#_Toc106657658)

[Fig 4.1 System Model 12](#_Toc106657659)

[Fig 5.2 Dataset sample 15](#_Toc106657660)

[Fig 5.3 Data after preprocessing 17](#_Toc106657661)

[Fig 5.4 Dataset Splitting Pie chart 19](#_Toc106657662)

[Fig 6.1 Accuracy comparison 25](#_Toc106657663)

[Fig 6. Precision and Recall Comparison 25](#_Toc106657664)

# Chapter 1

# INTRODUCTION

# 1.1 Machine learning

In machine learning, data plays a crucial role and perhaps the most important component.

Everything revolves around the data. The data is used to train algorithm for predicting, grouping, classifying, or searching for interesting patterns to reveal important insights. These important insights facilitate decision making in business and many other applications. Almost all fields are hit by this advancement and the future scope are quite enormous.

# 1.2 Types of machine learning

# 1.2.1 Supervised machine learning

Supervised learning needs output data. Part of the dataset is passed to the ML algorithm for training. This triggers the idea behind how the data actually behaves. During the training on the data, the ML algorithm detects the relationship between the specified parameters, thereby established the relation. The results are provided for the use in the final dataset. The algorithm is trained with new data will continue to improve after deployment.

# 1.2.2 Unsupervised Machine Learning

This algorithm doesn’t works need output data but requires a huge dataset to search for and create hidden structures. Hidden structures or relationships between data are recognized in an abstract way because there is no input. The ability of hidden structures makes unsupervised learning very versatile.

# 1.2.3 Reinforcement Learning

This algorithms use trial and error methods to self-improve and learn from new situations. Beneficial results are encouraged or "enhanced", and unfavorable results are discarded.

# 1.3 Application of Machine Learning

The most trending application of Machine Learning are:

1. Image Recognition
2. Automatic Language Translation
3. Speech Recognition
4. Medical Diagnosis
5. Stock Market Trading
6. Online Fraud Detection
7. E-mail Spam and Malware Filtering
8. Virtual Personal Assistant

## Chapter 2

## LITERATURE SURVEY

## 2.1 Natural Language Processing (NLP)

"NLP is a function of computer programs which will understand human language when speaking and writing."

It is one of branch of AI and has its roots in the field of linguistics. NLP uses artificial intelligence to take input which is in the form of written language or spoken, process the input data and interpret it in a way computer can understand.

## 2.2 NLP Tasks

Human language is full of ambiguity and it is incredibly hard to program that will give actual meaning of the text.

Some of the NLP tasks are:

1. Speech recognition converts speech into text. It is difficult due to the way people speak, such as different accents and the use of incorrect grammar.
2. PoS tagging.
3. Word-sense disambiguation selects word meanings using multiple meanings that determine the most meaningful word in a particular context.
4. Named entity extraction (NEM) identifies a word or phrase as a useful entity. Identify "Nirjuli" as the location and "George" as the male name.
5. Sentiment analysis seeks to extract subjective traits (attitude, emotion, irony, confusion, distrust) from the text.

## 2.3 Sentiment Analysis

In sentiment analysis, text are check for its sentiment i.e. positive, negative, or neutral. The sentiment analysis system combines NLP along with ML to score emotions for sentence or phrase entities, reviews, topics, and categories. The Sentiment Analysis is ideal for performing market research, polling, monitoring brand and product reputation, and more.

# 2.4 Types of sentiment analysis

1. Good mood indicates the type of polarity and can be assessed according to the problem.
2. Emotion recognition identifies emotions such as happiness, sadness, anger, and frustration.
3. Intent-based analytics not only captures opinions, but also the actions behind the text.
4. Aspect-based analysis captures certain positive or negative characteristics.

## 2.5 Sentiment Analysis Applications

• Sentiment analysis tools can be used by organizations in a variety of applications, including:

• Assessing brand awareness, reputation and level popularity over time or at a particular point in time. • Monitor consumer acceptance of new products or features.

• Evaluate the success of your marketing campaign.

• Define target audience or demographics.

• Collect customer feedback from social media, websites or online forms.

• Conduct market research.

• Categorize customer service requests.

## 2.6 Sentiment Analysis Challenges

1. Sentiment analysis-related tasks typically revolve around the inaccuracy of the training model

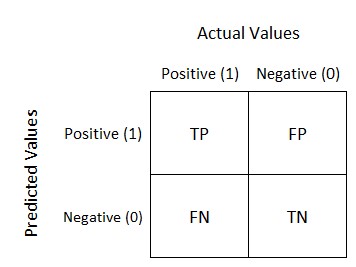
Comments with an objectivity or neutral tone tend to cause problems with the system and are often misunderstood.

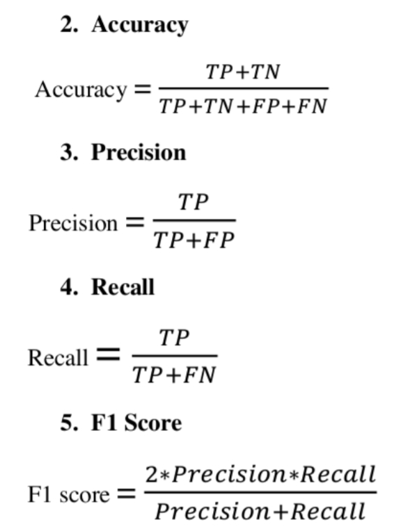
1. Emotions can also be difficult to identify if the system does not understand the context or tone
2. Computer programs also have problems with pictograms and irrelevant information. Particular attention should be paid to training models with emoji and neutral data to avoid mislabeling the text.
3. In many cases, people can be inconsistent in their remarks therefore it becomes more difficult for computers to analyze them.

## 2.7. Some basic terminology

**1. Confusion Matrix**

A table that will tell the performance of a model.



****

## 6. K-fold Cross Validation

It is an approached of splitting the dataset to optimized the effectiveness of dataset to build a more generalized model.

Overfitting is a common situation in training and prediction, where the model will predict the perfect score on trained data while the performance drop drastically on unseen data. To overcome this problem, dataset is split K-times, some part of the data will be used for training while the remaining for prediction thereby it validates the whole dataset K times.

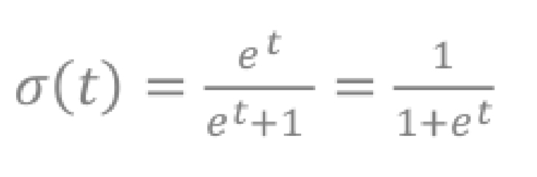
### Chapter 3

### METHOLOGY

### 3.1 Logistic Regression

#### It is a method that predicts the binary output.

The standard logistic function/ Sigmoid function is defined as follows:



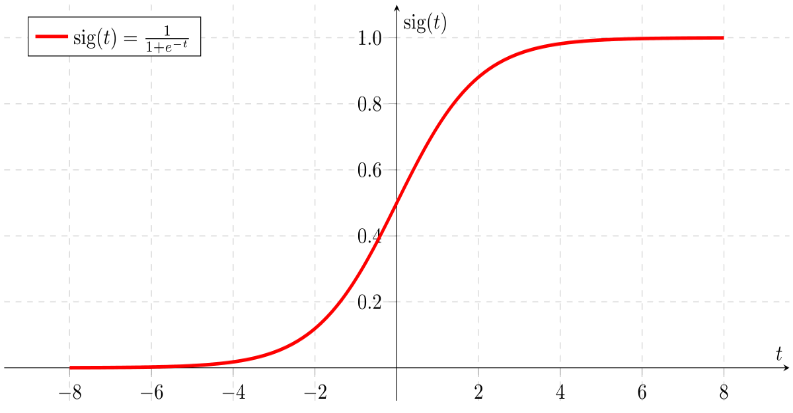
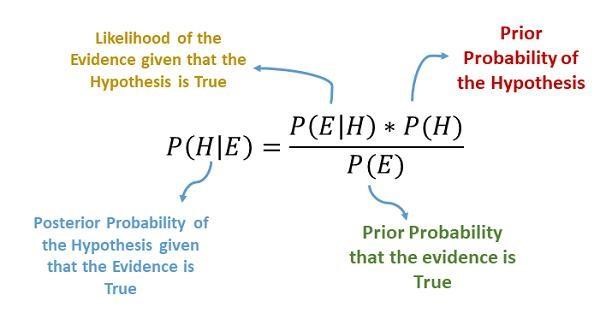


Fig 3.1 Standard logistic function (𝑡)

### 3.2 Naïve Bayes Model

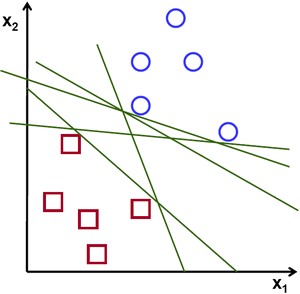
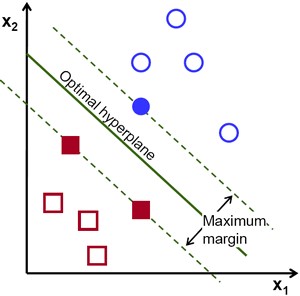
Naive Bayes model is based on Bayes' theorem, it assume all features that predict target values ​​are independent of each other. Calculate the probabilities for each class and select the one with the highest probability. Used in many applications and works well with Natural Language Processing (NLP).

The given is the Bayes’ Theorem relation.

Fig 3.2 .Bayes Theorem

### 3.3 Support Vector Machine

SVM will calculate the hyperplane in multi-dimensional space for data points seperation.

**Fig 3.3**. Possible hyperplanes

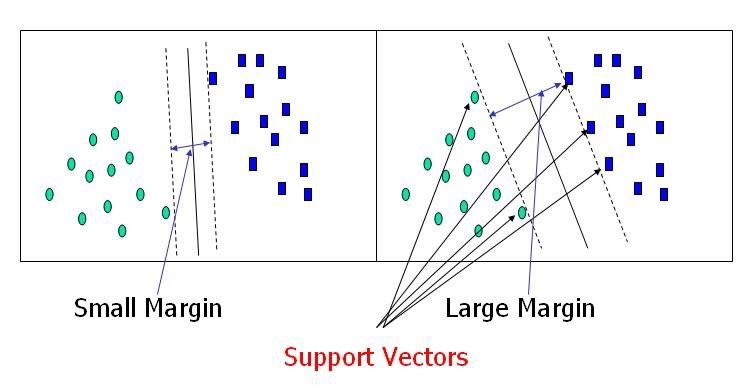
****

Fig 3.4. Support Vector Machine

### 3.4. Decision Tree

Decision Tree is based on hierarchical flow. The nodes represent the dataset character,branches as if else statement and final decision classification taken at the leaf node

The typical decision tree from decision node to leaf are shown.

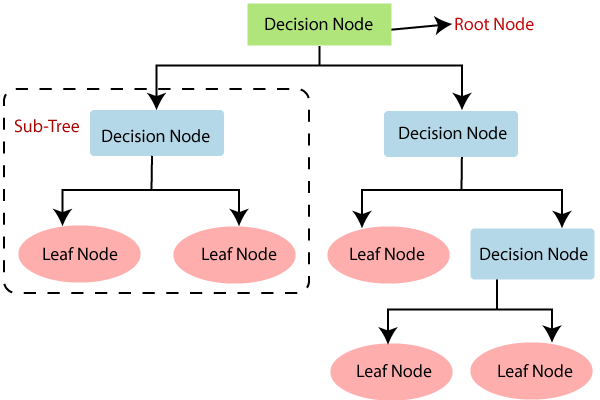


Fig 3.5. General structure of Decision Tree

### 3.5. Random Forest

Random Forest concept is based on combining multiple classifiers to building the model as well as improving the performance of the model.

It consist of multiple decision tree, each tree trained on original data sample. The final decision takes from the majority of vote giving by the trees.

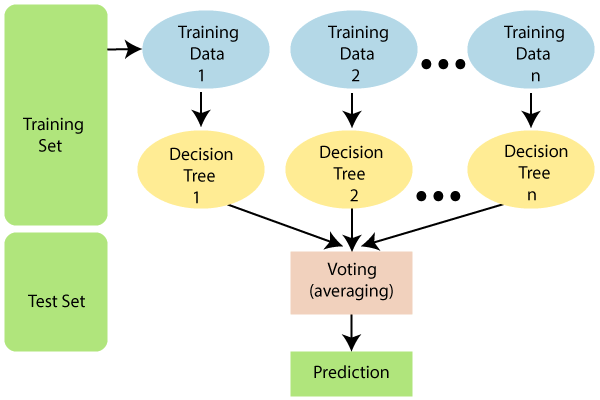


Fig 3.6 Random Forest Structure

### 3.6 Software Description

1. **Python**

Python 3.0 is used.

**2. NLTK**

NLTK is toolkit for NLP.

#### 3.7 Software requirements

OPERATING SYSTEM: Windows 10

IDE : Python 3.0

WEB APPLICATION : Jupyter notebook

INTERNET : Low Bandwidth is enough

BROWSER : Google Chrome

#### Chapter 4

#### DESIGN

#### 4.1 Design Flow

To implement an algorithm for classification of review into Positive or Negative with highest accuracy possible. The review can be gathered to summarize the overall sentiment on a particular topic.

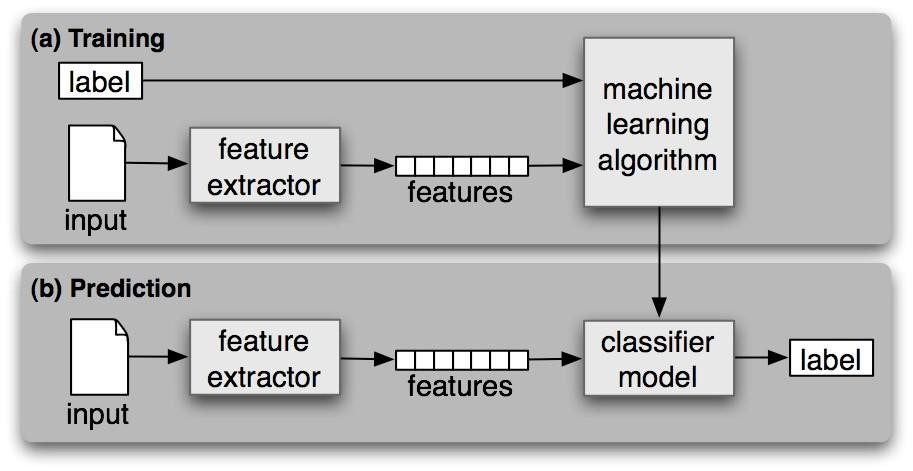


Fig 4.1 System Model

#### 

#### 4.1.2 Training data

Training data consist of review given by the customer and its sentiment

#### 4.1.3. Features

All the symbols, stop words, words without any sentiment are removed. The remaining list of words remain in the review after preprocessing are features, which have some sentiments.

#### 4.1.3. Model

It creates trained classifier. In our case we used five algorithm to train namely Naïve Bayes algorithm, Logistic Regression, SVM and random forest.

#### 4.1.5. Input

The inputs are the reviews (sentences/words)

#### 4.1.6. Label

Sentiment output generated by the classifier where output of 0 indicates negative review and output 1 indicates positive review

##### Chapter 5

##### IMPLEMENTATION

##### 5.1 Implementation flowchart

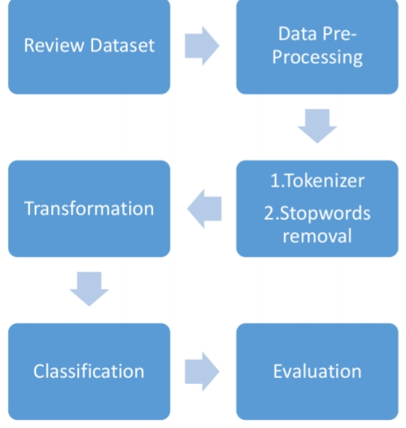


Fig 5.1. Flowchart

##### 5.2. Review Dataset

The datasets Restaurant\_Reviews has imported which is basically a tsv file. It has thousand reviews, some sample of reviews are shown below.

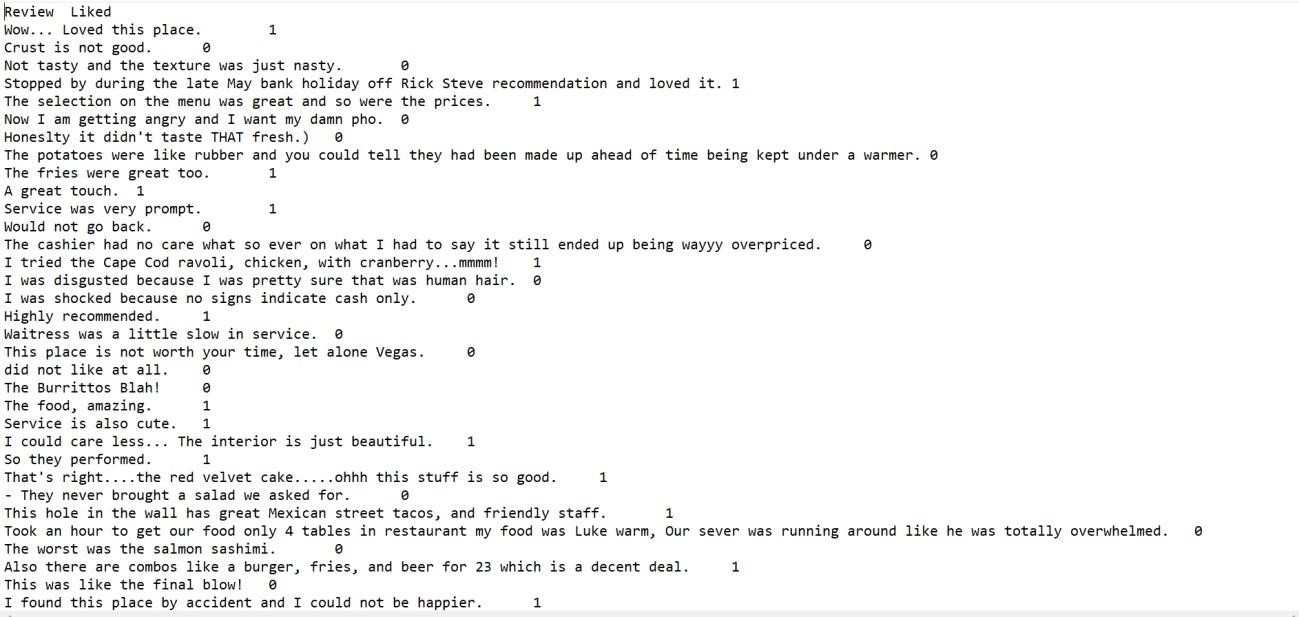


Fig 5.2 Dataset sample

##### 5.3 Data Preprocessing

##### 5.3.1. Punctuation

Punctuation marks such as the comma, semicolon, full stop, quotation marks, and ellipsis points, among others, are frequently redundant because they offer no value or meaning to the NLP model. As a result, the model removes any punctuation that appears in the reviews. For example, **wow...the food, amazing** changes to **wow the food amazing**

##### 5.3.2. Tokenization

It's the technique of separating strings into word lists. That splitting is done using Regular Expressions. For example, **Service was very prompt** becomes ['Service', 'was', 'very', ‘prompt’].

##### 5.3.3. Stop words

Eliminating irrelevant words.

##### 5.3.4 Stemming

It's the method of reducing a word to its simplest form. The major goal is to reduce the number of different spellings of the same term, reducing the number of words in the model.

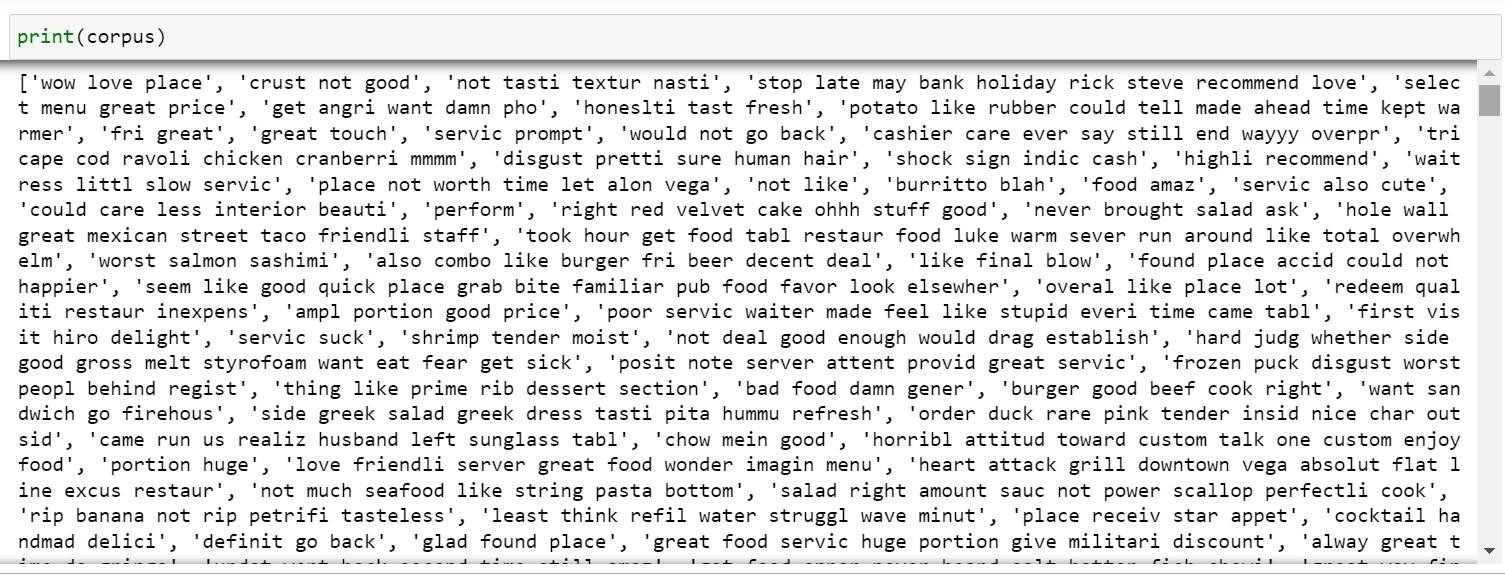


Fig 5.3 Data after preprocessing

##### 5.4. Transformation

##### 5.4.1 Creating Bags of words

Algorithm usually works on numbers in NLP. The text cannot be feed directly to the algorithm. Therefore it needs transformation to process further. This is where Bag of words comes in. Bags of words is used to preprocess the text converting it into a bag of words, which count occurrences of most frequently used words.

Example:

Review 1: ‘Wow loved place’

Review 2: ‘Crush burger loved’

|  |  |
| --- | --- |
|  | **Frequency** |
| Wow | 1 |
| Loved | 2 |
| Crush | 0 |
| Place | 1 |
| Burger | 0 |

Review 1 frequency in the vector form

Review 1- [1, 2, 0, 1, 0]

|  |  |
| --- | --- |
| **Words** | **Frequency** |
| Wow | 0 |
| Loved | 2 |
| Crush | 1 |
| Place | 0 |
| Burger | 1 |

Review 2 in the vector form

Review 2- [0, 2, 1, 0, 1]

##### 5.5. Splitting of dataset into Training and Test set

Of the thousand reviews we got from the customers, 80% of the data will be going to training set and the remaining 20% will be reserve for the test set.

Data set Percentage

Training set

Test set

Fig 5.4 Dataset Splitting Pie chart

##### 5.6. Training of Dataset

The review data set is trained by using five (5) different classification algorithm namely Naïve Bayes model, SVM, random forest, logistic regression, decision tree.

##### 5.6.1.Logistic Regression



##### 5.6.2. Naïve bayes

****

##### 5.6.3. SVM

****

##### 5.6.4. Decision tree

**** 5.6.5. Random forest

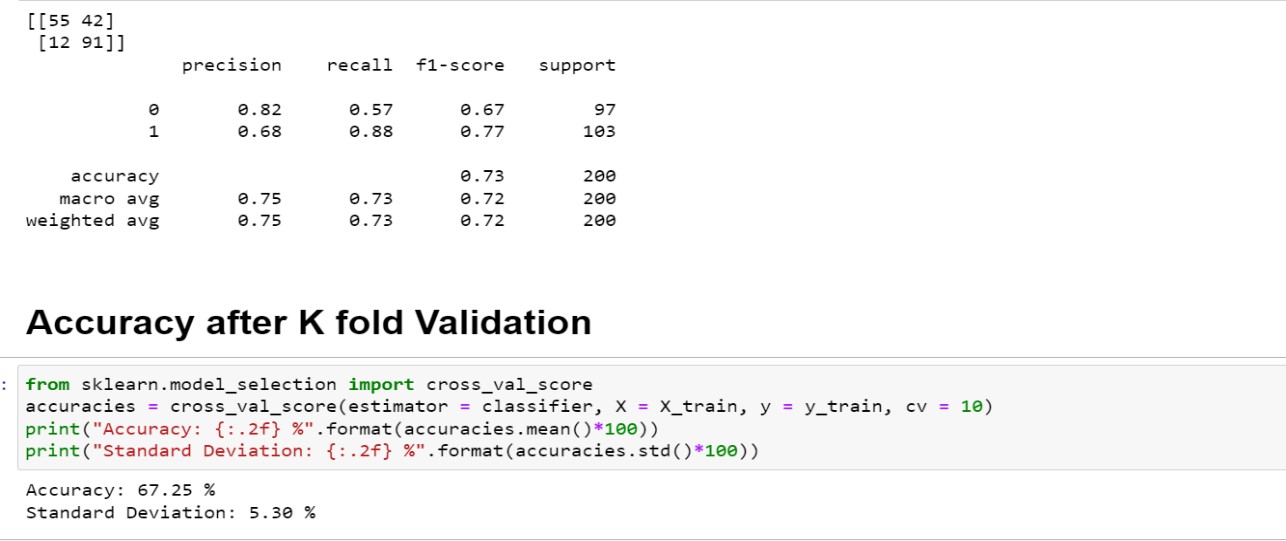
****

###### Chapter 6

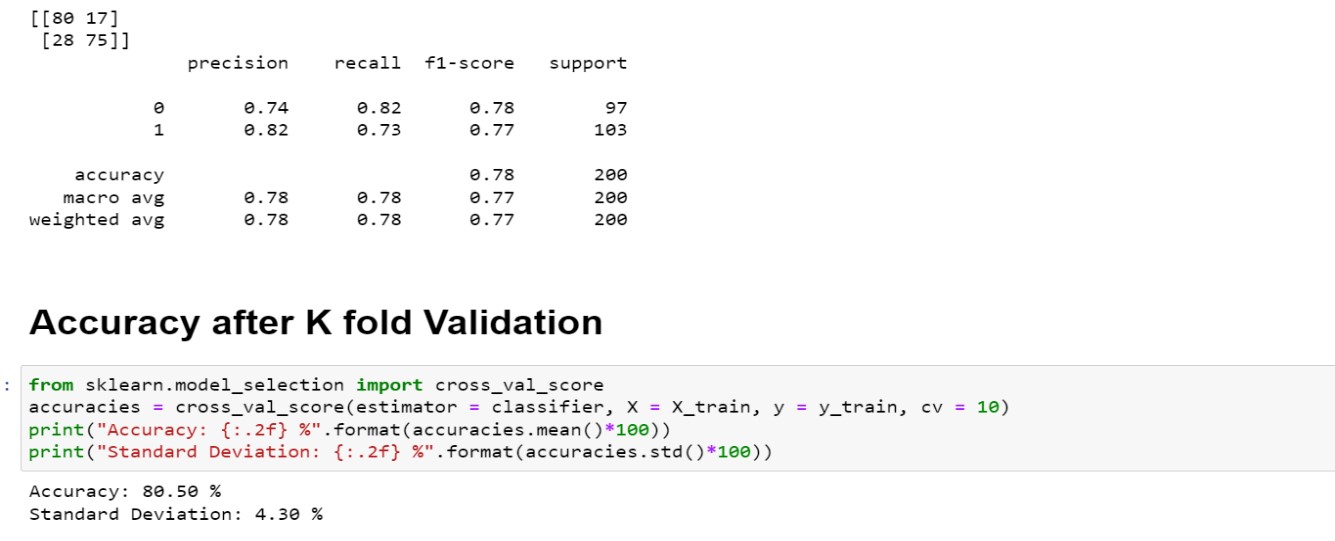
###### RESULT AND ANALYSIS

###### 6.1. Results based on Confusion Matrix

###### 6.1.1. Naïve Bayes Model

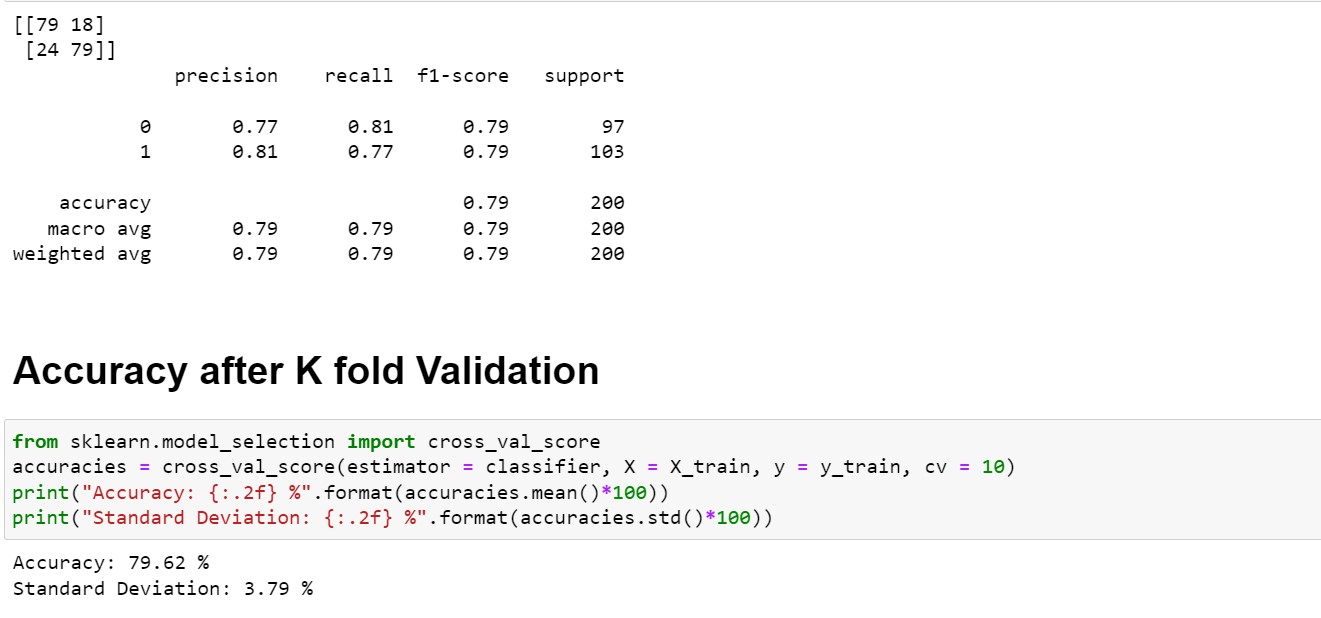
****

**6.1.2. Logistic Regression**

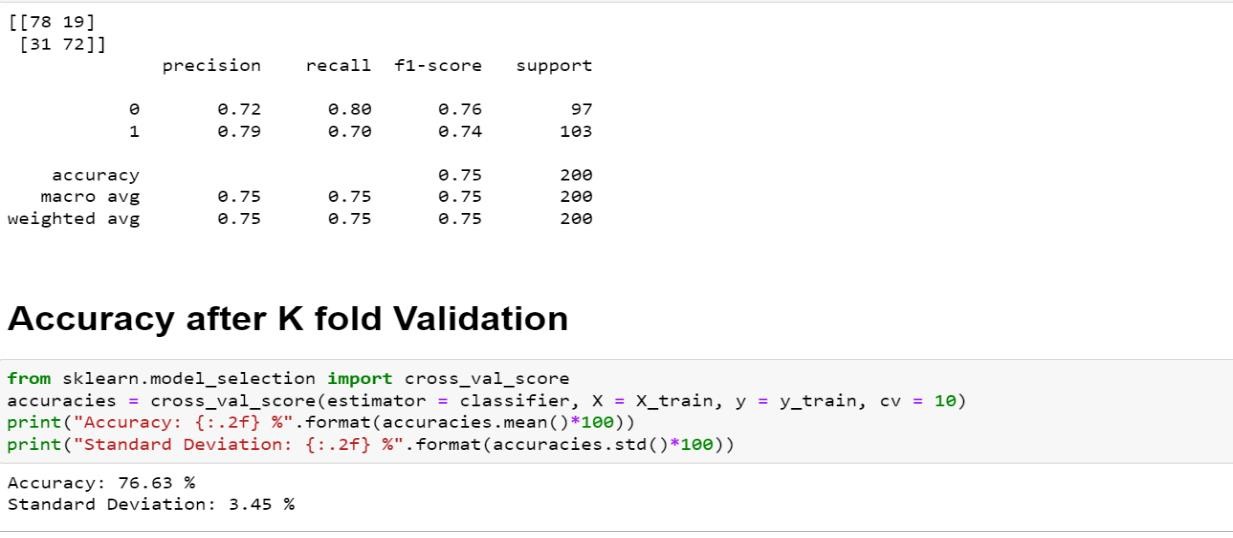


###### 6.1.3. Random forest

###### 6.1.4. Support Vector Machine

****

###### 6.1.5. Decision Tree

****

###### 6.2. Accuracy comparison between Models before and after Cross validation

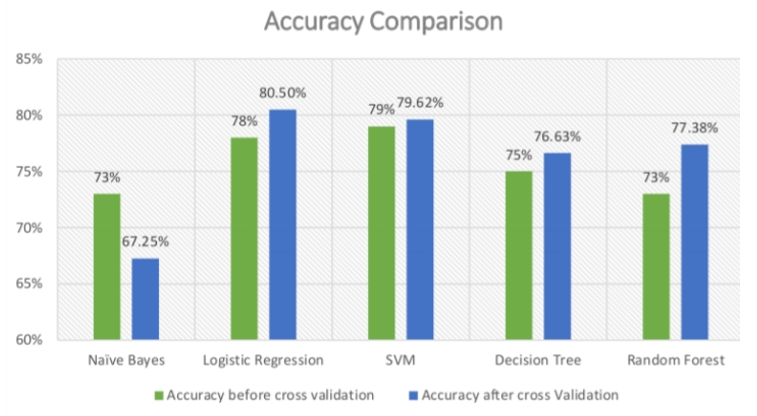


Fig 6.1 Accuracy comparison

###### 

###### 6.3. Precision and Recall comparison between the models

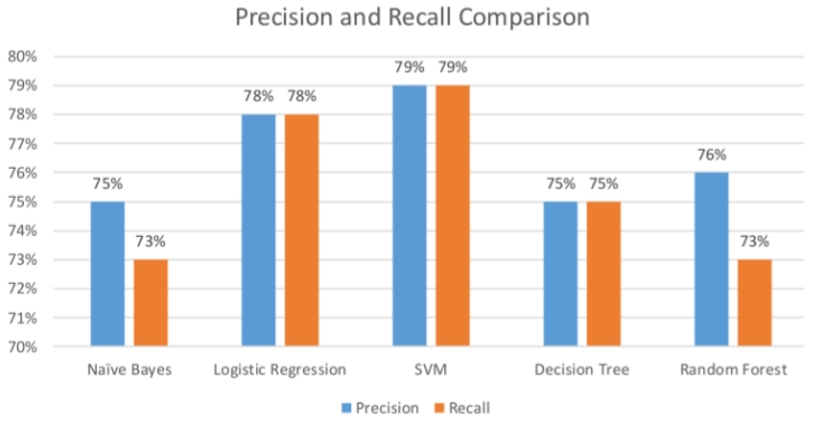


Fig 6. Precision and Recall Comparison

###### 6.4. Analysis

The review dataset is trained by using five different model. While evaluating the performance based on the confusion matrix, Logistic Regression has the highest accuracy after cross validation closely followed by SVM. Decision Tree and Random Forest performs almost equal with the accuracy while Naive Bayes perform the worst.

Based on the Precision and Recall, SVM edge Logistic Regression by small margin, the remaining order remains the same as decision tree, random forest and naïve bayes comes third, fourth and fifth respectively.

Therefore based on all the existing parameter from the confusion matrix, Logistic Regression is the best suited model for our Review dataset, SVM comes second. Decision Tree at third, at fourth Random Forest while the least performing model is the Naïve Bayes model.

We took the existing algorithm and tried to enhance the accuracy with different preprocessing technique and especially using K-Fold cross validation. It greatly optimize the performance of the model.

Chapter 7

CONCLUSION AND FUTURE SCOPE

Although sentiment analysis is beneficial, it is a difficult task. The difficulty rises in tandem with the intricacy of the conveyed viewpoints. Because of its effectiveness, sentiment analysis is in high demand. Thousands of text documents can be analyzed for sentiment in seconds, rather than the hours it would take a team of humans to accomplish it manually. Many firms are adopting text and sentiment analysis and incorporating it into their business processes since it is so efficient, accurate, and quick. Sentiment mining is the study of big groups of people and trends. It means you can allow for some fuzziness in sentiment categorization with the sheer amount of data; otherwise, we'll learn that the patterns we're looking for aren't really popular or significant.

Despite the positive results that can be achieved and the promise shown, there are still many challenges due to the difficulty of implementation due to the unclear natural language problem. In our case, the analysis of customer opinions given is a classic example, as they are often coupled with difficulty in determining the sentiment expressed. In addition, a large data set is required when emotional words are paired with sentimental values. Scanning is usually done in English, where there is a limitation for other languages

8. REFERENCES

1. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams engineering journal* 5.4 (2014)
2. Feldman, Ronen. "Techniques and applications for sentiment analysis." *Communications of the ACM* 56.4 (2013)
3. Hussein, Doaa Mohey El-Din Mohamed. "A survey on sentiment analysis challenges." *Journal of King Saud University-Engineering Sciences* 30.4 (2018)
4. Mejova, Yelena. "Sentiment analysis: An overview." *University of Iowa,*

*Computer Science Department* (2009).

1. Cambria, Erik, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco, eds.

"A practical guide to sentiment analysis." (2017)

1. Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." *Journal of the American Medical Informatics Association* 18.5 (2011): 544-551.
2. Grosz, Barbara J., Karen Sparck-Jones, and Bonnie Lynn Webber, eds. *Readings in natural language processing*. Morgan Kaufmann Publishers Inc., 1986.
3. Mahesh, Batta. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR).[Internet]* 9 (2020): 381-386.
4. Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma. "A review of supervised machine learning algorithms." *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee, 2016.
5. <https://www.lexalytics.com/technology/sentiment-analysis/>
6. <https://monkeylearn.com/sentiment-analysis/>
7. <https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications/>
8. <https://www.ibm.com/in-en/cloud/learn/machine-learning/>
9. <https://www.javatpoint.com/classification-algorithm-in-machine-learning/>
10. <https://www.ibm.com/cloud/learn/natural-language-processing/>
11. <https://scikit-learn.org/stable/supervised_learning.html/>
12. <https://www.kaggle.com/datasets/jurk06/restaurant-dataset>
13. <https://www.geeksforgeeks.org/understanding-logistic-regression/>
14. https://en.wikipedia.org/wiki/Natural\_language\_processing/

