### Twitter Sentiment Analysis



**BACHELOR OF ENGINEERING**

#### IN INFORMATION TECHNOLOGY

**By**

**Riya Kalburgi (Roll no. 24)**

**Punit Solanki (Roll no. 72)**

**Rounak Suthar (Roll no. 73)**

**Under the Guidance of Mrs. Deepali Patil**

Assistant Professor, Department of Information Technology



**Near Commissioner's Bungalow, Kanakia Park, Mira Road (E), Thane-401107, Maharashtra.**

### Department of Information Technology 2020-21

#### A Mini Project Report On

**Twitter Sentiment Analysis**

*Submitted to Mumbai University*

***In partial fulfillment for the award of the degree of***

### BACHELOR OF ENGINEERING

#### IN INFORMATION TECHNOLOGY

**By**

**Riya Kalburgi (Roll no. 24)**

**Punit Solanki (Roll no. 72)**

**Rounak Suthar (Roll no. 73)**

**Under The Guidance Of**

**Mrs. Deepali Patil**

Assistant Professor, Department of Information Technology



**Near Commissioner's Bungalow, Kanakia Park, Mira Road (E), Thane-401107, Maharashtra.**

### Department of Information Technology 2020-21

#### UNIVERSITY OF MUMBAI

CERTIFICATE

This is to certify that the project titled “Twitter Sentiment Analysis” has been completed under our supervision and guidance by the following students:

**Riya Kalburgi**

**Punit Solanki**

**Rounak Suthar**

In the partial fulfillment of degree of Bachelor of Engineering in Information Technology branch as prescribed by the University of Mumbai during the academic year 2020-2021. The said work has been assessed and is found to be satisfactory.

##### Signature of the Internal Examiner Name: Prof. Deepali Patil

**Date:**

**Signature of the External Examiner Name:**

**Date:**

**Signature of the H.O.D. Name: Prof. Sunil Yadav Date:**

**Signature of the Principal Name: Dr. S. Ram Reddy Date:**

**DECLARATION**

We do hereby declare that the work embodied in the project entitled “**Twitter Sentiment Analysis”** is the outcome of our original work under the guidance and supervision of **Prof. Deepali Patil**. This piece of work or any part of it has not been submitted previously for the award of any other degree, diploma, or other title to any other institution.

We also declare that this written submission represents our ideas in our own words and where others ideas or words have been included. We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: May, 2021

**Riya Kalburgi**

Roll No.: 24

Exam. Seat No.:

**Punit Solanki**

Roll No.: 72 Exam. Seat No.:

**Rounak Suthar**

Roll No.: 73 Exam. Seat No.:

**ACKNOWLEDGEMENT**

A few sublime human experiences defy expressions of any kind, and a feeling of true gratitude is one of them. I, therefore, find words quite inadequate to express my indebtedness to my Guide **Prof. Deepali Patil** for their virtuous guidance, encouragement and help throughout this work. Their deep insight into the problem and the ability to provide solutions has been immense value in improving the quality of project at all stages. This experience of working with them shall ever remain a source of inspiration and encouragement for me.

I express my thanks to **Prof. Sunil Yadav**, HOD (IT), SLRTCE, Mira Road, for extending his support that he gave truly help the progression of the project work.

My sincere thanks to **Dr. S. Ram Reddy**, Principal, SLRTCE, Mira Road for providing me the necessary administrative assistance in the completion of the work.

I am extremely grateful to the celebrated authors whose precious works have been consulted and referred in my project work. I also wish to convey my appreciation to my friends who provided encouragement and timely support in the hour of need.

Special thanks to my Parents whose love and affectionate blessings have been a constant source of inspiration in making this a reality.

All the thanks are, however, only fraction of what is due to Almighty for granting me an opportunity and the divine grace to successfully accomplish this assignment.

##### Riya Kalburgi

##### Punit Solanki

##### Rounak Suthar

|  |  |  |
| --- | --- | --- |
| Chapter No | Topic | Page No |
| 1 | INTRODUCTION |  |
|  | 1.1 Description | …11 |
|  | 1.2 Problem Formulation | …11 |
|  | 1.3 Motivation | …12 |
|  | 1.4 Proposed Solution | …12 |
|  | 1.5 Scope of Project | …13 |
| 2 | REVIEW OF LITERATURE |  |
|  | 2.1 Literature Survey | …15 |
|  | 2.2 Problem Statement | …15 |
| 3 | Description   * 1. Theory      1. Data preprocessing      2. Naïve Bayes Algorithm      3. Natural Language Processing      4. Performance measurement – Explain Confusion matrix   2. Design Considerations   3.2.1 Overview of system (Draw the flowchart of your project) |  |
|  |  |
|  | …17 |
|  | …19 |
|  | …22 |
|  | …23  …24 |
| 4 | Implementation and results |  |
|  | 4.1 Naïve Bayes Algorithm - Steps of algorithm with screenshot of output  4.2 Real Time Analysis of NLP- Steps of algorithm with screenshot of output | …27  …33 |
|  | 4.3 Dataset- Explanation of dataset that you are using | …40 |
| 5 | Comparative study |  |
|  | 5.1 Comparison of two algorithm-graphically | …42 |
|  |  |  |
| 6 | Conclusion  6.1 Conclusion  6.2 Future scope | …45  …45 |
| 7 | References | …46 |

**ABSTRACT**

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative, neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. Several Machine Learning algorithms can used to perform opinion mining on data. This project leverages the concepts of Naive Bayes algorithm and Natural Language Processing to achieve the results. Furthermore, the result obtained from each of these algorithms are compared to understand their respective suitability under varied conditions.

Keywords: Naive Bayes Algorithm, Natural Language Processing, Twitter Sentiment Analysis, R programming.

### LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| Fig. No | Name | Page No |
| 1 | Flowchart for Naïve Bayes Algorithm | 24 |
| 2 | Flowchart for Natural Language Processing | 25 |
| 3 | Sentiment in Sample Text using NLP | 38 |
| 4 | Shiny Application User Interface | 39 |
| 5 | Sentiment Analysis of Hollywood actor Dwayne Johnson | 39 |
| 6 | Sentiment Analysis of Indian cricketer Sachin Tendulkar | 40 |
| 7 | Dataset | 41 |
| 8 | Accuracy of pattern analysis (Naïve Bayes) | 43 |
| 9 | Accuracy of pattern analysis (NLP) | 43 |

**Chapter 1**

**Introduction**

**Chapter 1: Introduction**

This chapter will introduce the reader with Twitter Sentiment Analysis System. It will light up the topics like the description of the project and the former formulation of the problem behind it as well as what motivated the makers of the project to take a decision to make this project and its related problem solution and thus covering up the scope of the project.

**1.1 Description**

An idea or view based on emotion is often referred to as a sentiment. Analyzing these emotional opinions from different groups of users and determining their attitude towards the complete contextual polarity or emotional response to a document, communication, or event is done in sentimental analysis. Sentiment analysis allows user to know the circumstance of what the review are made regarding specific product or ongoing matters rather than understanding deep thoughts of the reviewer. The text is analyzed for determining sentiment of the person or group of people and classified into positive, negative, and neutral. Twitter, serve as one of the most accepted platforms to convey the opinions and views of a particular product or topic. If people find topics relevant or interesting, then they would desire to share their opinion about the topic. The topic could be a product or a service or a social message or any other object. Understanding this can help us decide the type of posts the user needs to increase such things for more user engagement through various ways. Analysis of social media facts like tweets helps people to get an idea of a particular product or topic which makes it easy to take decisions on it. Sentiment analysis also helps people to change their attitude about wrong belief on a product, service, or a topic. It helps people to choose which is best by analyzing the comments or tweets of a particular topic or a product. For e.g.: During elections people go through the reviews to make it simple to decide or get an idea on whom to choose or elect the right candidate in the elections.

**1.2 Problem Formulation**

These days, technology has got its new and higher pace. This development has changed human’s way of expressing their opinions, sentiments and views and the platforms in which they do so. Micro blogging websites are just social networking webpage on which people write regular and short posts. One of the most famous micro blogging services is twitter where people can post and read messages which can be 148 characters long. Today around approx. 6500 tweets are tweeted every second, which roughly brings out 561.6 million tweets for every day. These streams of tweets are mainly noisy reflecting multi topic, changing states of mind information in unfiltered and unstructured format. Analyzing unstructured data is in itself a difficult task and extracting useful information from it’s a big challenge. For doing this, there is a need of powerful tools and technologies which can help to handle millions of tweets and extracting sentiment from them. There are many different possible ways to do this. R is an open-source approach used for analyzing on-line reviews to perform sentiment analysis and text mining. Sentimental Analysis is a strategy to explore whether a gathered content is in positive, negative, or neutral state. Essentially, it involves examining the emotions related with a piece of writing for any topic. Sentiment analysis is used to check the opinions, taste, views and interest of individuals by seeing diverse prospective, for example, celebrity, politicians, foods, places, or some other topic. In sentimental analysis we usually classify everyone’s mood in various classifications.

**1.3 Motivation**

Comments, reviews, and opinion of the people play an important role to determine whether a given population is satisfied with the product, services. It helps in predicting the sentiment of a wide variety of people on a particular event of interest like the review of a movie, their opinion on various topic roaming around the world. These data are essential for sentiment analysis. In order to discover the overall sentiment of population, retrieval of data from sources like Twitter, Facebook, Blogs are essential. For the sentiment analysis, we focus our attention on the Twitter, a micro-blogging social networking website. Twitter generates huge data that cannot be handled manually to extract some useful information and therefore, the ingredients of automatic classification are required to handle those data. Tweets are unambiguous short texts messages that are up to a maximum of 140 characters. By the use of Twitter, millions of people around the world to be connected with their family, friends and colleagues through their computers or mobile phones. The Twitter interface allows the user to post short messages and that can be read by any other Twitter user. Twitter contains a variety of text posts and grows every day. We choose Twitter as the source for opinion mining simply because of its popularity and data mining.

**1.4 Proposed Solution**

Social networking sites acquired immense popularity and interest with the people around the world. Twitter is one of the effective tools for any business intelligence to get information about what people are talking and reacting about the topics that are roaming around the world. A twitter helps to engage the users and directly communicates with them and in response, users to provide word-of-mouth marketing for companies by discussing the product quality. With the limited resources and knowing about no one can target directly to the destination consumers, the business intelligence can be more efficient in their policy of marketing by being very selective about consumers choice they should reach out to. The proposed solution is to perform sentiment analysis on Twitter data. Sentiment Analysis can be performed using a number of Machine Learning Algorithms. This project leverages the concepts of Naive Bayes algorithm and Natural Language Processing to achieve the results. Furthermore, the result obtained from each of these algorithms are compared to understand their respective suitability under varied conditions.

**1.5 Scope**

Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange.

Sentiment Analysis Dataset Twitter has a number of applications:

**Business**: Companies use Twitter Sentiment Analysis to develop their business strategies, to assess customers’ feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.

**Politics**: In politics Sentiment Analysis Dataset Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset Twitter is also used for analysing election results.

**Public Actions**: Twitter Sentiment Analysis also is used for monitoring and analysing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.

# Chapter 2

**Review of Literature**

**Chapter 2: Review of Literature**

Here we will elaborate the aspects like the literature survey of the project and what all projects are existing and been used in the market which the makers of this project took the inspiration from and thus decided to go ahead with the project covering with the problem statement. Literature review helps us analyze the past innovations related to the project and also help ameliorate them.

**2.1 Literature Survey**

[1] Twitter Data in the form of opinion, feedback, reviews, remarks and complaint are treated as big data and it cannot be used directly. These data first convert as per requirement. In this paper, pre-processing of data to remove noise from the data is discussed. It has implemented sentiment analysis for movie data set, on Hadoop framework using Naïve Bayes algorithm and analyzed with large number of tweets. This type of analysis will definitely help any organization to improve their business productivity. The analysis of twitter data is done on various perspective like Positive, Negative and Neutral sentiments on tweets. It also provides the fast-downloading approach for efficient Twitter Trend Analysis. Tweets can also be useful in prediction of product sales, quality of services offered by company, feedback of users etc. Hence, the future scope in the sentiment analysis for the other social networking websites like Facebook, Google Plus etc.

[2] This paper developed a model to sentiment analysis which allows the processing of Twitter API streaming feed in real time and to classify its polarity to provide valuable insight in industry and users. This paper has used Natural Language Processing to achieve the desired results. The built classifier can be utilized as data analysis tools in NLTK. Therefore, in general the proposal technique can be used for sentiment analysis for any device, public figure or sports team that is better than any other existing model with high accuracy performance.

**2.2 Problem Statement**

The online social media such as Twitter, Facebook, and Instagram allow users to communicate with the whole world. Write their own opinions about products or share their moments, even influence politics and companies. Twitter for example, almost every huge company have an account on Twitter to know about their customers feedback about their services or products. Sentiment analysis, known as opinion mining, for classifying specific words into positive or negative.

In this paper, we used sentiment analysis to classify specific English tweets using two algorithms namely, Naïve Bayes algorithm and Natural Language Processing. Our research was determining which one better than other, in specific we examined weather specific tweets is positive, negative, neutral.

# Chapter 3

**Description**

**Chapter 3: Description**

This chapter enlightens about the theoretical explanation of the concepts used in the implementation of the project such as data preprocessing, algorithms used, and performance measure followed by the design consideration details.

**3.1 Theory**

Sentiment Analysis is a technique widely used in text mining. Twitter Sentiment Analysis, therefore, means, using advanced text mining techniques to analyse the sentiment of the text (here, tweet) in the form of positive, negative, and neutral. It is also known as Opinion Mining, is primarily for analysing conversations, opinions, and sharing of views (all in the form of tweets) for deciding business strategy, political analysis, and for assessing public actions.

Enginuity, Revealed Context, Steamcrab, MeaningCloud, and SocialMention are some of the well-known tools used for the analysis of Twitter sentiment. R and Python are widely used for [sentiment analysis](https://www.digitalvidya.com/blog/sentiment-analysis/) dataset twitter. Sentiment Analysis of Twitter data is now much more than a college project or a certification program. A good number of Tutorials related to Twitter sentiment are available for educating students on the Twitter sentiment analysis project report and its usage with R and Python.

**3.1.1 Data Pre-processing**

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

**Steps Involved in Data Pre-processing:**

**1. Data Cleaning:**  
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

**(a). Missing Data:**  
This situation arises when some data is missing in the data. It can be handled in various ways.  
Some of them are:

**1. Ignore the tuples:**  
This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

**2. Fill the Missing values:**  
There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

**(b). Noisy Data:**  
Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

**1. Binning Method:**  
This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

**2. Regression:**  
Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

**3. Clustering:**  
This approach groups the similar data in a cluster. The outliers may be undetected, or it will fall outside the clusters.

**2. Data Transformation:**  
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**  
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**  
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**  
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**  
   Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**  
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**  
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**  
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**  
   This enables to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**  
   This reduces the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

**3.1.2 Naïve Bayes Algorithm**

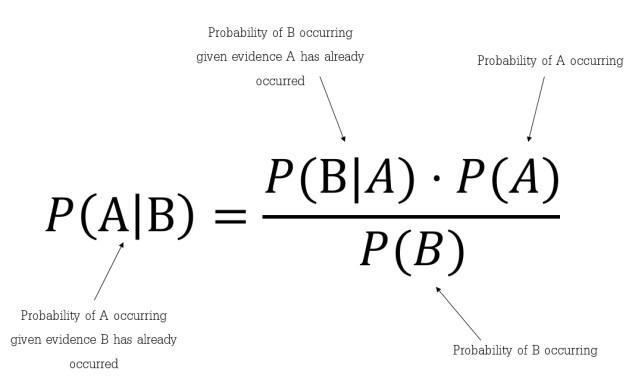
The simplest solutions are usually the most powerful ones, and Naïve Bayes is a good example of that. Despite the advances in Machine Learning in the last years, it has proven to not only be simple but also fast, accurate, and reliable. It has been successfully used for many purposes, but it works particularly well with natural language processing (NLP) problems.

Naïve Bayes is a probabilistic machine learning algorithm based on the **Bayes Theorem**, used in a wide variety of classification tasks. In this article, we will understand the Naïve Bayes algorithm and all essential concepts so that there is no room for doubts in understanding.

### Bayes Theorem Bayes’ Theorem is a simple mathematical formula used for calculating conditional probabilities.

**Conditional probability** is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred.

The formula is: —

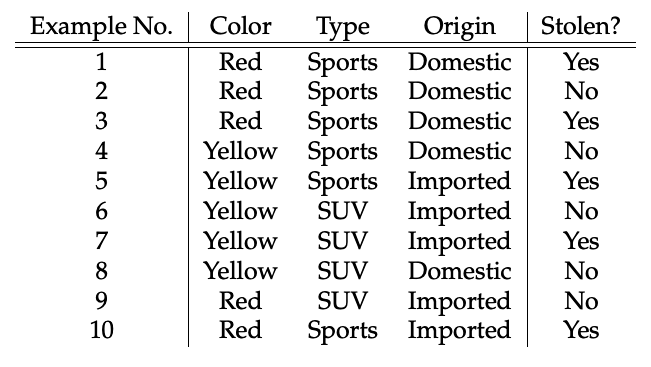


Which tells us: how often A happens given that B happens, written **P(A|B)**also called posterior probability, when we know: how often B happens given that A happens, written **P(B|A)** and how likely A is on its own, written **P(A)** and how likely B is on its own, written **P(B).**

In simpler terms, Bayes’ Theorem is a way of finding a probability when we know certain other probabilities.

### Example

The dataset is represented as below.



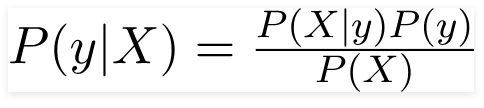
Concerning our dataset, the concept of assumptions made by the algorithm can be understood as:

* We assume that no pair of features are dependent. For example, the color being ‘Red’ has nothing to do with the Type or the Origin of the car. Hence, the features are assumed to be **Independent**.
* Secondly, each feature is given the same influence (or importance). For example, knowing the only Color and Type alone can’t predict the outcome perfectly. So, none of the attributes are irrelevant and assumed to be contributing **Equally** to the outcome.

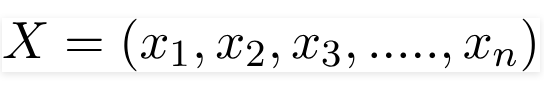
**Note:** The assumptions made by Naïve Bayes are generally not correct in real-world situations. The independence assumption is never correct but often works well in practice. **Hence the name ‘Na**ï**ve’.**

Here in our dataset, **we need to classify whether the car is stolen, given the features of the car**. The columns represent these features, and the rows represent individual entries. If we take the first row of the dataset, we can observe that the car is stolen if the Color is Red, the Type is Sports and Origin is Domestic. So we want to classify a Red Domestic SUV is getting stolen or not. Note that there is no example of a Red Domestic SUV in our data set.

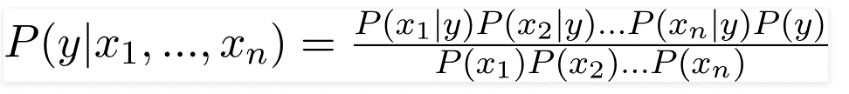
According to this example, Bayes theorem can be rewritten as:



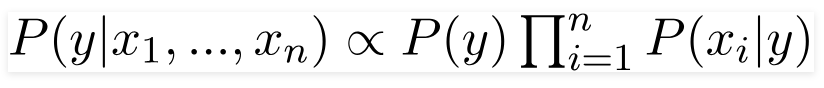
The variable **y** is the class variable(stolen?), which represents if the car is stolen or not given the conditions. Variable **X**represents the parameters/features.  
**X** is given as,



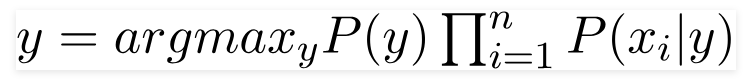
Here x1,x2….xn represent the features, i.e they can be mapped to Color, Type, and Origin. By substituting for **X**and expanding using the chain rule we get,



Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remains static. Therefore, the denominator can be removed, and proportionality can be injected.

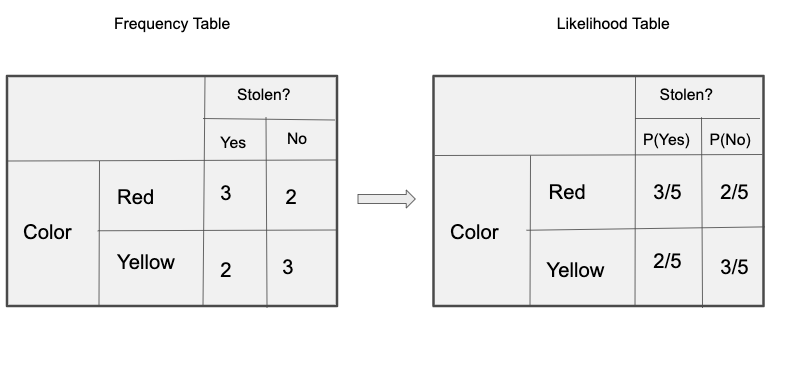


In our case, the class variable(**y**) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we have to find the class variable(**y)** with maximum probability.

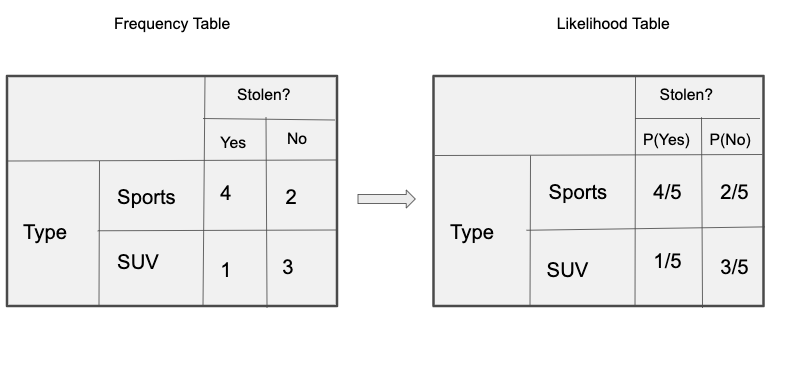


Using the above function, we can obtain the class, given the predictors/features.

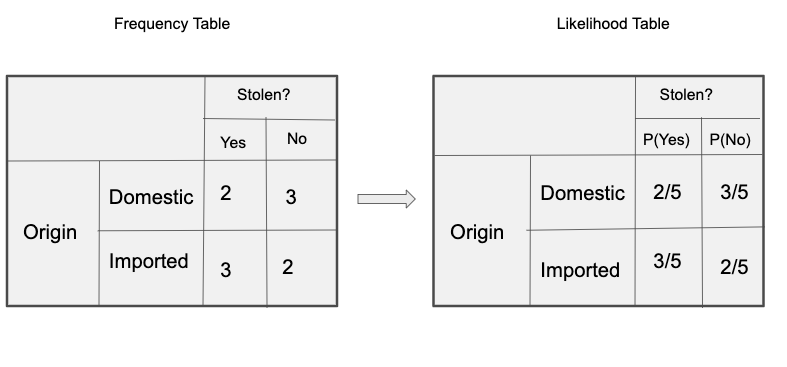
The posterior probability **P(y|X)** can be calculated by first, creating a **Frequency Table** for each attribute against the target. Then, molding the frequency tables to **Likelihood Tables** and finally, use the Naïve Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction. Below are the Frequency and likelihood tables for all three predictors.



Frequency and Likelihood tables of ‘Color’

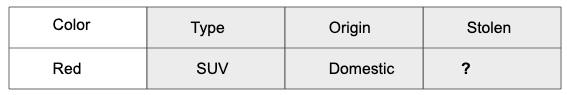


Frequency and Likelihood tables of ‘Type’

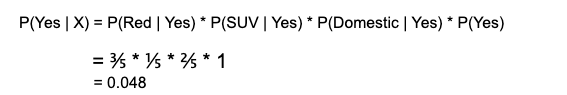


Frequency and Likelihood tables of ‘Origin’

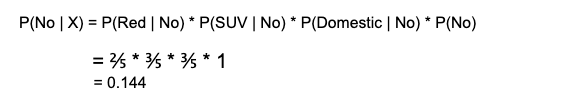
So in our example, we have 3 predictors **X**.



As per the equations discussed above, we can calculate the posterior probability P(Yes | X) as :



and, P(No | X):



Since 0.144 > 0.048, Which means given the features RED SUV and Domestic, our example gets classified as ’NO’ the car is not stolen.

**3.1.3 Natural Language Processing**

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.

Most NLP techniques rely on machine learning to derive meaning from human languages.

NLP entails applying algorithms to identify and extract the natural language rules such that the unstructured [language data](https://blog.liveedu.tv/a-quick-introduction-to-text-summarization-in-machine-learning/) is converted into a form that computers can understand.

When the text has been provided, the computer will utilize algorithms to extract meaning associated with every sentence and collect the essential data from them.

Sometimes, the computer may fail to understand the meaning of a sentence well, leading to obscure results.

For example, a humorous incident occurred in the 1950s during the translation of some words between the English and the Russian languages.

Here is the biblical sentence that required translation:

“The spirit is willing, but the flesh is weak.”

Here is the result when the sentence was translated to Russian and back to English:

“The vodka is good, but the meat is rotten.”

**3.1.4 Performance Measure (Confusion Matrix)**

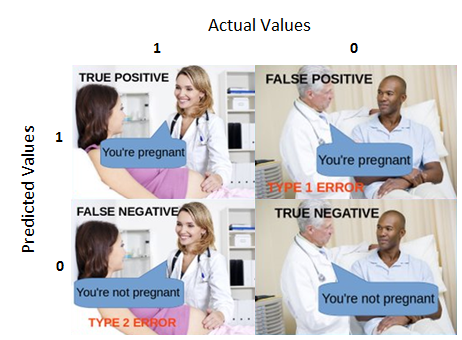
A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



It is extremely useful for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

Let’s understand TP, FP, FN, TN in terms of pregnancy analogy.



**True Positive:**

Interpretation: You predicted positive and it’s true.

You predicted that a woman is pregnant, and she actually is.

**True Negative:**

Interpretation: You predicted negative and it’s true.

You predicted that a man is not pregnant, and he actually is not.

**False Positive: (Type 1 Error)**

Interpretation: You predicted positive and it’s false.

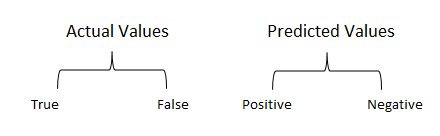
You predicted that a man is pregnant, but he actually is not.

**False Negative: (Type 2 Error)**

Interpretation: You predicted negative and it’s false.

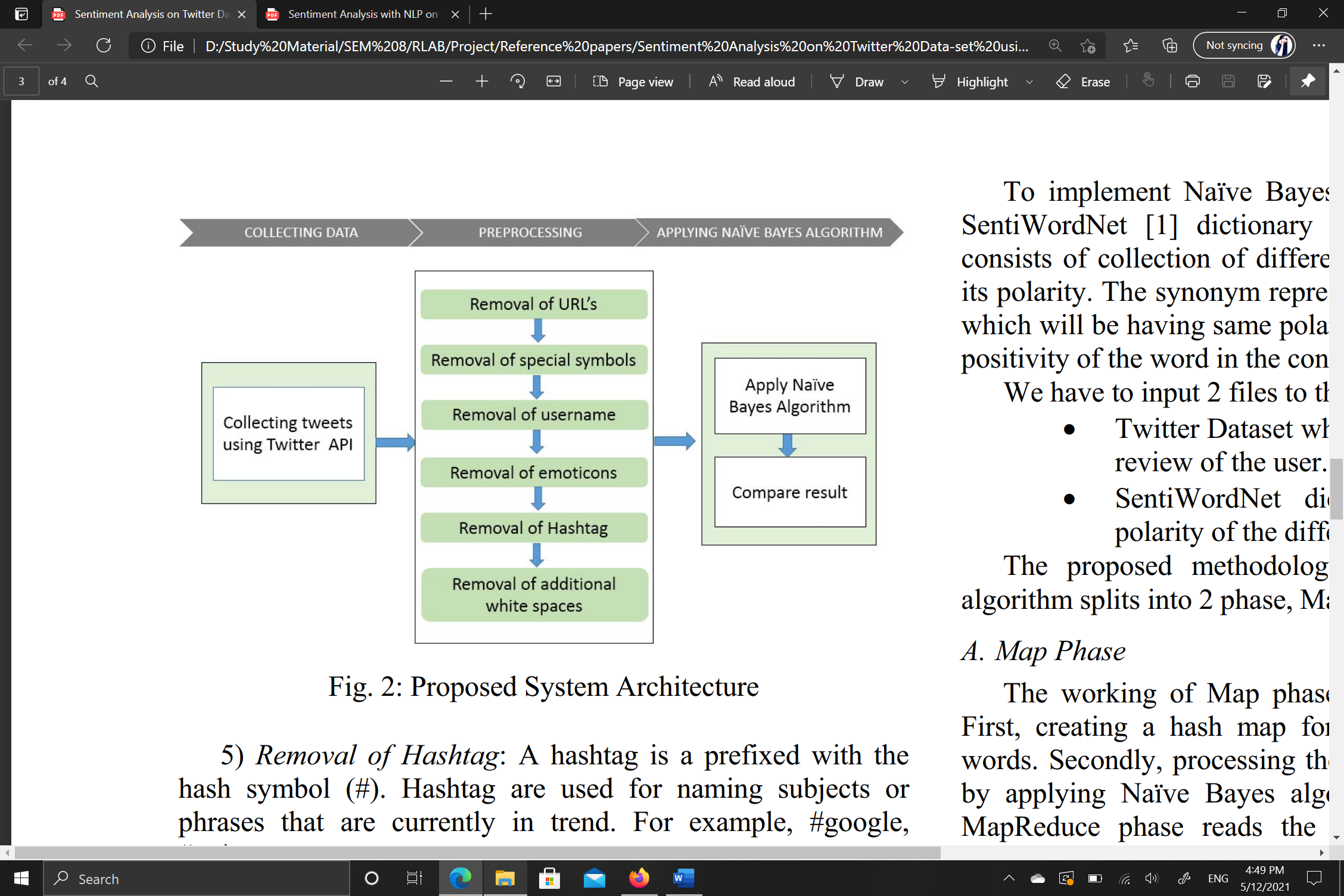
You predicted that a woman is not pregnant, but she actually is.

Just Remember, we describe predicted values as Positive and Negative and actual values as True and False.

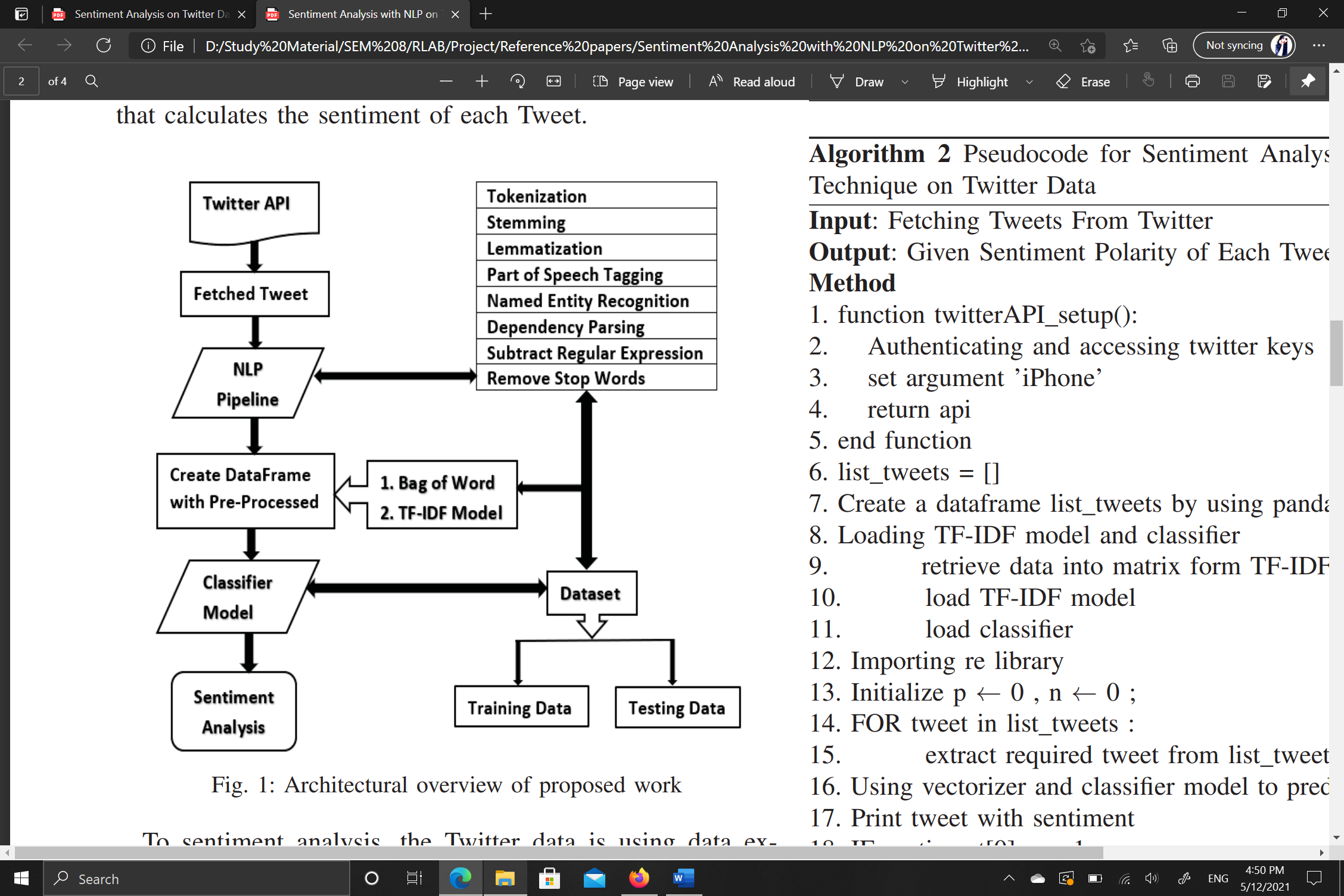


**3.2 Design Considerations**

**3.2.1 Overview of System**



*Figure 1: Flowchart for Naïve Bayes Algorithm*



*Figure 2: Flowchart for NLP*

# Chapter 4

**Implementation and results**

**Chapter 4: Implementation and Results**

In this chapter the Implementation details of the project would be shed light upon including the lines of code executed to achieve the results.

**4.1 Naïve Bayes Algorithm**

**4.1.1 Installing and importing all the necessary packages**

#A machine learning package for automatic text classification that makes it simple for novice #users to get started with machine learning, while allowing experienced users to easily #experiment with different settings and algorithm combinations.

**>library(RTextTools)**

#R supports a package called ‘e1071’ which provides the naive bayes training function.

**>library(e1071)**

#The caret package contains functions to streamline the model training process for complex #regression and classification problems.

**>library(caret)**

#word cloud generator package helping us to analyze texts and to quickly visualize the #keywords as a word cloud.

**>library(wordcloud)**



**>set.seed(12345)**

**4.1.2 Providing dataset for testing and training purpose**

#Given dataset of positive sentiment tweets

**>sentPositive <- c(**

**"I like it", "like it a lot", "It's really good",**

**"recommend!", "Enjoyed!", "like it",**

**"It's really good", "recommend too",**

**"outstanding", "good", "recommend!",**

**"like it a lot", "really good",**

**"Definitely recommend!", "It is fun",**

**"liked!", "highly recommend this",**

**"fantastic show", "exciting",**

**"Very good", "it's ok",**

**"exciting show", "amazing performance",**

**"it is great!","I am excited a lot",**

**"it is terrific", "Definitely good one",**

**"very satisfied", "Glad we went",**

**"Once again outstanding!", "awesome"**

**)**

# dataset of negative sentiment tweets

**>sentNegative <- c(**

**"Not good at all!", "rude",**

**"It is rude", "I don't like this type",**

**"poor", "Boring", "Not good!",**

**"not liked", "I hate this type of",**

**"not recommend", "not satisfied",**

**"not enjoyed", "Not recommend this.",**

**"disgusting movie","waste of time",**

**"feel tired after watching this",**

**"horrible performance", "not so good",**

**"so boring I fell asleep", "poor show",**

**"a bit strange","terrible"**

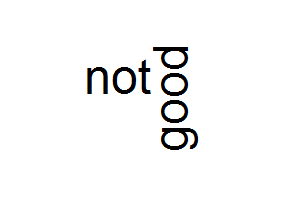
**)**

#Generating keywords from dataset

**>wordcloud(sentPositive, max.words = 40, scale = c(3, 0.5))**



**>wordcloud(sentNegative, max.words = 40, scale = c(3, 0.5))**



**4.1.3 Splitting data into training and testing set**

#creating a single dataframe to hold both positive and negative tweets

**>df = data.frame(sentiment="positive", text=sentPositive)**

**>df = rbind(df, data.frame(sentiment="negative", text=sentNegative))**

# creating a sample of ramdow tweets from df to split data in testing and training set

**>index = sample(1:nrow(df), size = .9 \* nrow(df))**

# 90% of data from dataset used for training

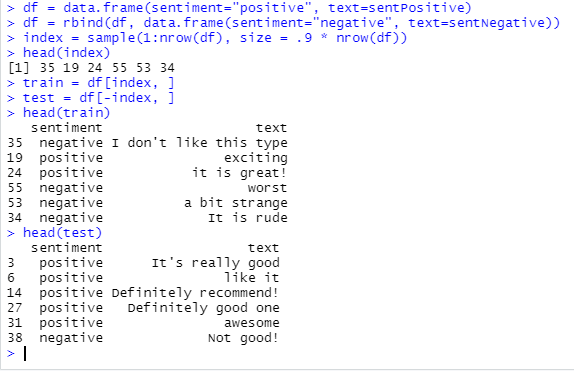
**>train = df[index, ]**

#remaining 10% which is not included in train set is used for testing

**>test = df[-index, ]**

**>head(train)**

**>head(test)**



**4.1.4 Preparing document matrix from the text of a train and test data with a create\_matrix function of the RTextTool package**.

**>mTrain = create\_matrix(train[,2], language = "english",**

**removeStopwords=FALSE, removeNumbers=TRUE,**

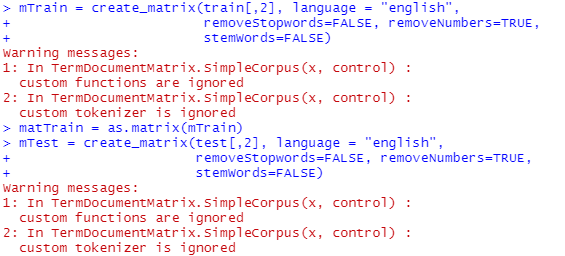
**stemWords=FALSE)**

**>matTrain = as.matrix(mTrain)**

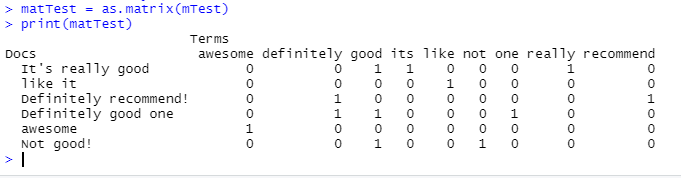
**>mTest = create\_matrix(test[,2], language = "english",**

**removeStopwords=FALSE, removeNumbers=TRUE,**

**stemWords=FALSE)**



**>matTest = as.matrix(mTest)**

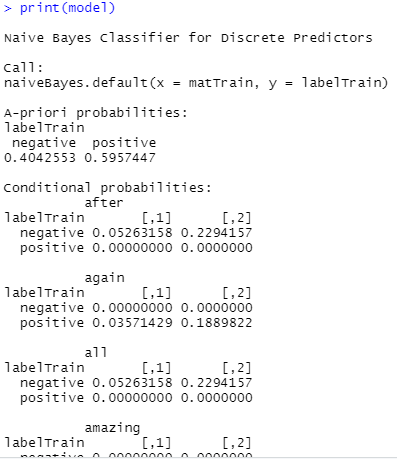


**>labelTrain = as.factor(train[,1])**

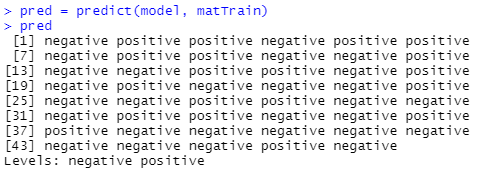
**>labelTest = as.factor(test[,1])**

**4.1.5 Creating Model by using NaïveBayes function from “e1071” package which provides training and testing functionalities.**

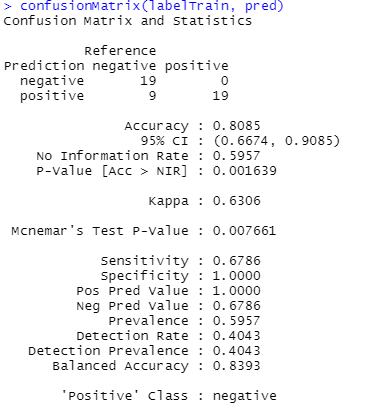
**>model = naiveBayes(matTrain, labelTrain)**



**>pred = predict(model, matTrain)**

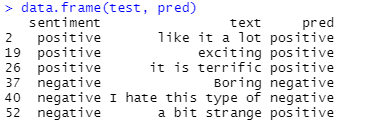


**confusionMatrix(labelTrain, pred)**

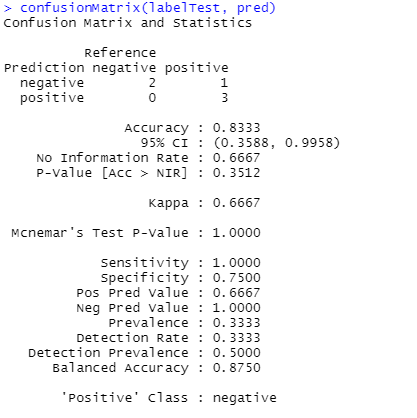


**>pred = predict(model, matTest)**

**>data.frame(test, pred)**



**>confusionMatrix(labelTest, pred)**



**4.2 Real Time Analysis using NLP**

**4.2.1 Installing and loading necessary libraries**

#shiny package provides GUI support to R project

**>library("shiny")**

#An R interface to the C 'libstemmer' library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary

**>library("SnowballC")**

#Provides an interface to the Twitter web API.

**>library("twitteR")**

#The package comes with four sentiment dictionaries and provides a method for accessing #the robust, but computationally expensive, sentiment extraction tool developed in the NLP #group at Stanford.

**>library("syuzhet")**

**4.2.2 Creating User Interface for shinyApp**

**>sidebar\_content <- sidebarPanel(**

**textInput("y\_var",label = "user\_id"),**

**actionButton("update", "Change")**

**)**

**>main\_content <- mainPanel(**

**plotOutput("plot")**

**)**

**>panel <- tabPanel(**

**"Visualization",**

**titlePanel("Whose sentiment analysis you want to do?"),**

**p("Please enter twitter user\_id."),**

**sidebarLayout(**

**sidebar\_content, main\_content**

**)**

**)**

**# User Interface -----------------------------------------------------**

**>ui <- navbarPage(**

**"Sentiment analysis",**

**panel**

**)**

**#--------------------------------------------------------------------------------------------------------**

**4.2.3 Creating server for app with logic code**

**>server <- function(input, output, session) {**

**name <- ""**

# Update data and rerun app only if change button is clicked

**observeEvent(input$update, {**

#credentials to access Twitter API

**consumer\_key <- 'dFQd5EoJbywMHBCEax8PTdwVu'**

**consumer\_secret <- 'zBTK3jJlKzP6fQ7ZO0fxKnURldwHgz08TsgwYdTmOfClOdJMvz'**

**access\_token <- '1380533717081280512-OGSMrTfVmpiC4MHoJ71slV8WPKtJkB'**

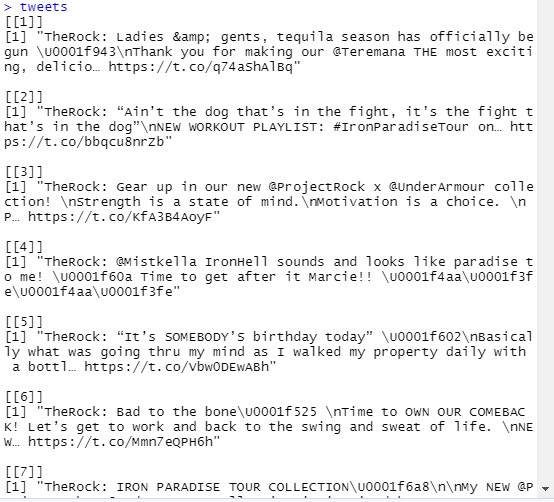
**access\_secret <- 'vXDAdM2EnEQ3HT5zWx8pAdUnmbne6DvwFclyge5O4mt7r'**

**setup\_twitter\_oauth(consumer\_key, consumer\_secret, access\_token, access\_secret)**

**name <- input$y\_var**

#fetch latest 200 tweets of given twitter user

**tweets <- userTimeline(name, n=200)**



**n.tweet <- length(tweets)**

**tweets.df <- twListToDF(tweets)**

# Cleaning the dataset

**tweets.df2 = gsub("(RT|via)((?:\\b\\W\*@\\w+)+)", "", tweets.df$text);**

**tweets.df2 = gsub("@\\w+", "", tweets.df2); # regex for removing @user**

**tweets.df2 = gsub("[[:punct:]]", "", tweets.df2); # regex for removing punctuation mark**

**tweets.df2 = gsub("[[:digit:]]", "", tweets.df2); # regex for removing numbers**

**tweets.df2 = gsub("http\\w+", "", tweets.df2);# regex for removing links**

**tweets.df2 = gsub("#", " ", tweets.df2);**

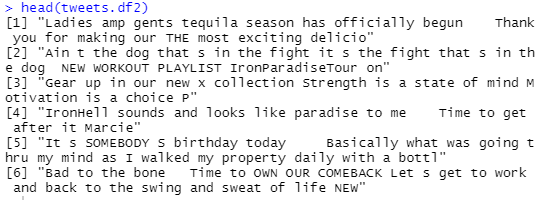
**tweets.df2 = gsub("\n", " ", tweets.df2); ## regex for removing new line (\n)**

**tweets.df2 = gsub("[ \t]{2,}", " ", tweets.df2); ## regex for removing two blank space**

**tweets.df2 = gsub("[^[:alnum:]///' ]", " ", tweets.df2) # keep only alpha numeric**

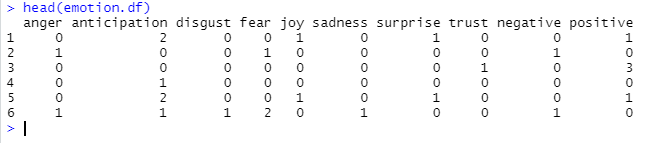
**tweets.df2 = iconv(tweets.df2, "latin1", "ASCII", sub="") # Keep only ASCII characters**

**tweets.df2 = gsub("^\\s+|\\s+$", "", tweets.df2);**

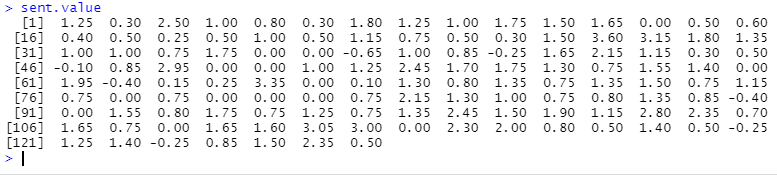


**word.df <- as.vector(tweets.df2)**

**emotion.df <- get\_nrc\_sentiment(word.df)**



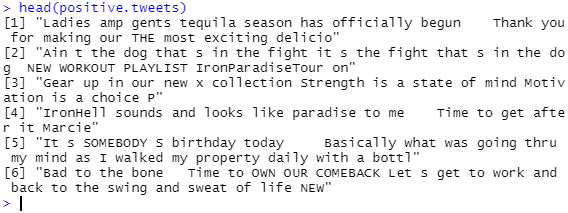
**sent.value <- get\_sentiment(word.df)**



**most.positive <- word.df[sent.value == max(sent.value)]**



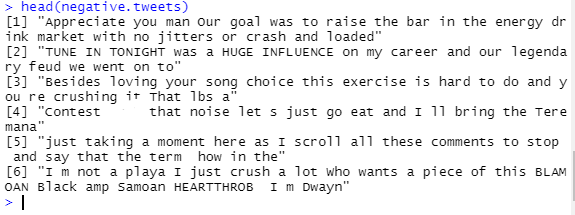
**positive.tweets <- word.df[sent.value > 0]**



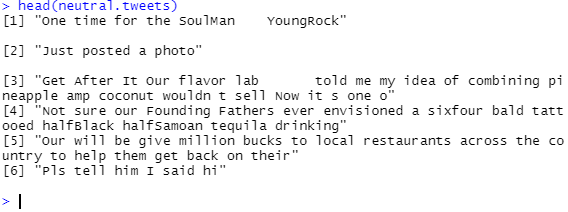
**most.negative <- word.df[sent.value <= min(sent.value)]**



**negative.tweets <- word.df[sent.value < 0]**



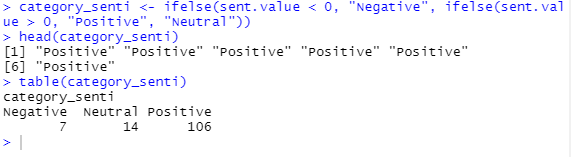
**neutral.tweets <- word.df[sent.value == 0]**



**category\_senti <- ifelse(sent.value < 0, "Negative", ifelse(sent.value > 0, "Positive", "Neutral"))**

**head(category\_senti)**

**table(category\_senti)**



# Finally creating a plot on the percentage of sentiments analyzed.

**output$plot <- renderPlot({**

**barplot(**

**sort(colSums(prop.table(emotion.df[, 1:8]))),**

**horiz = TRUE,**

**cex.names = 0.7,**

**las = 1,**

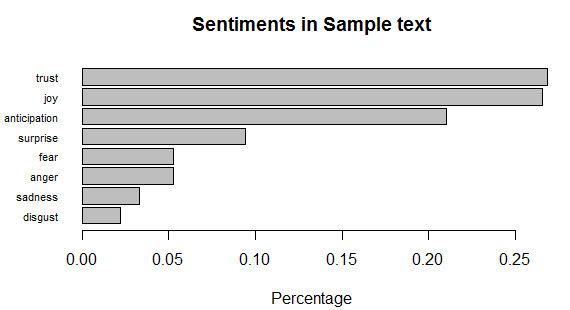
**main = "** **Sentiments in Sample text", xlab="Percentage"**

**)**

**})**

**})**

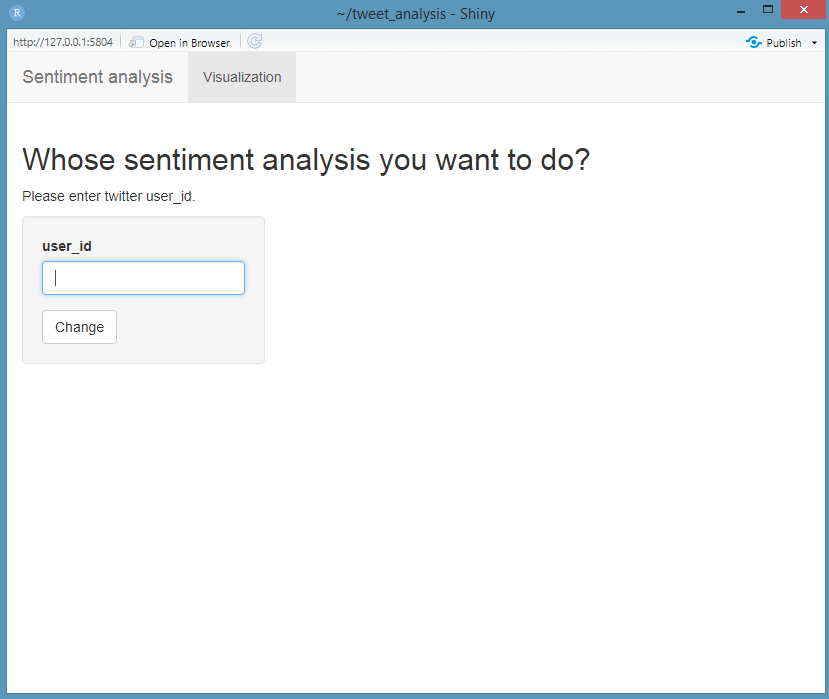
**}**



*Figure 3: Sentiment in Sample Text using NLP*

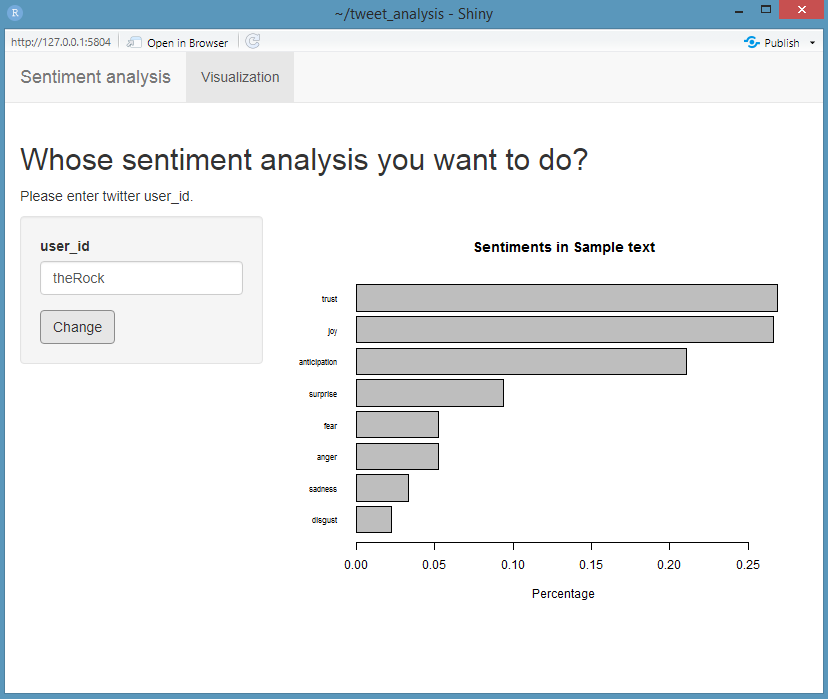
# Command to run shinyApp

**shinyApp(ui,server)**



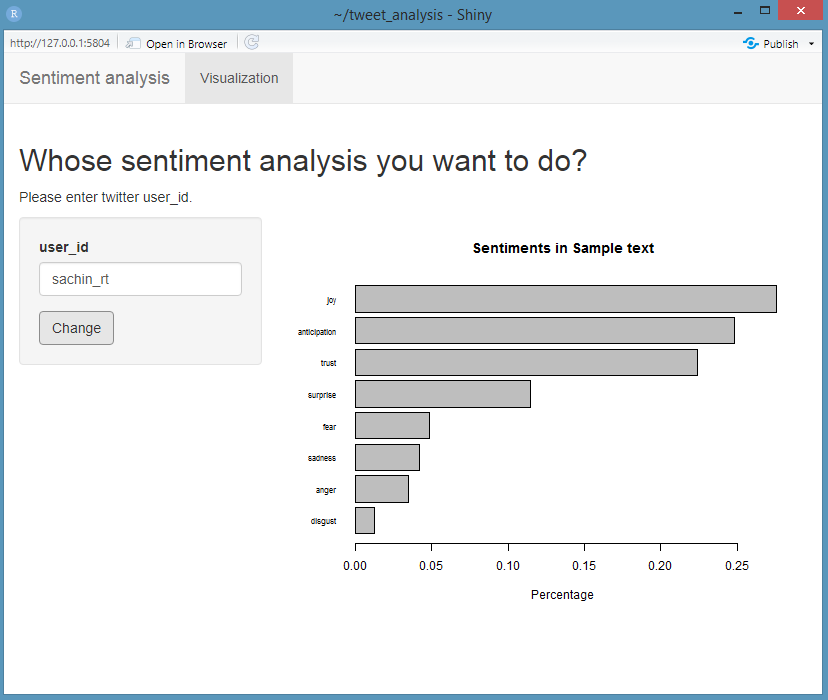
*Figure 4: Shiny Application User Interface*

#Tweet sentiment analysis of Hollywood actor Dwayne Johnson



*Figure 5: Sentiment Analysis of Hollywood actor Dwayne Johnson*

# Tweet sentiment analysis of Indian cricketer Sachin Tendulkar

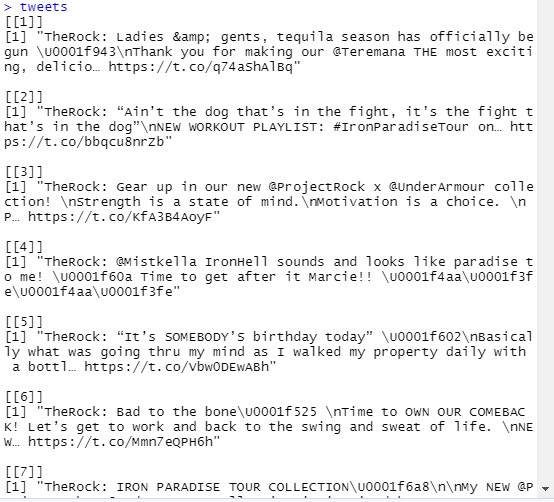


*Figure 6: Sentiment Analysis of Indian cricketer Sachin Tendulkar*

**4.3 Dataset Explanation**

The algorithm operates on real-time Twitter Data. This is achieved by using Twitter API. **API for** streaming **that provides access to a high volume of tweets** with low latency. Basic features of the Twitter *API*

* The Twitter API has four main “objects”: Tweets, Users, Entities and Places.
* It has daily restrictions for calls and changes in the API to protect Twitter from abuses. Specifically, the restriction is set up by the user, or better said, by a user access token. The frequency restrictions are divided into 15-minute intervals and all the evaluation criteria require authentication so unauthenticated calls cannot be made to the API.
* The API is based on HTTP (over SSL), so the processes that require a specific HTTP method will return an error if the request is not correct.
* There are specific parameters for requests to the API, generated paging and library restrictions to adapt API operation to this social network.



*Figure 7: Dataset*

# Chapter 5

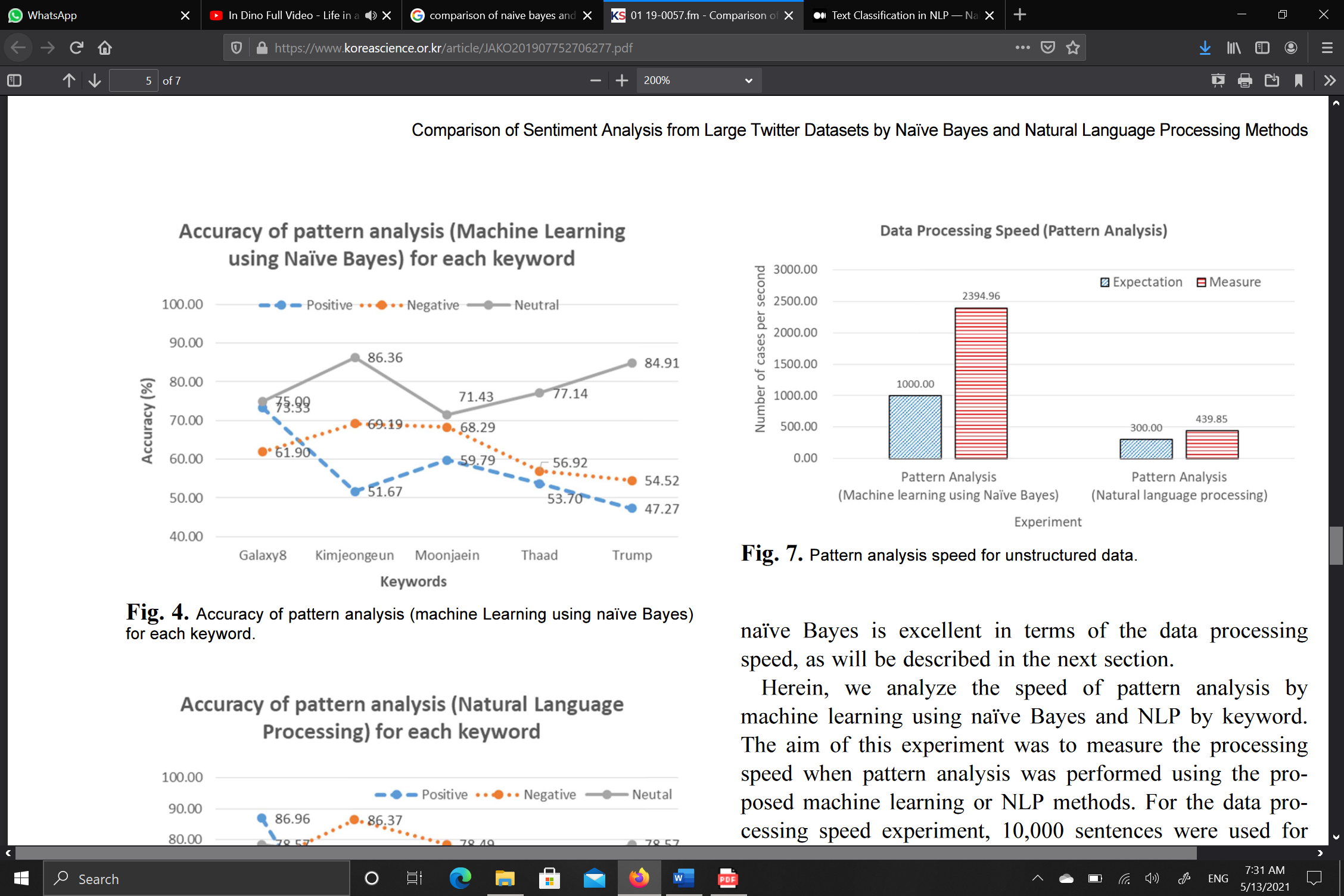
**Comparative study**

**Chapter 5: Comparative Study**

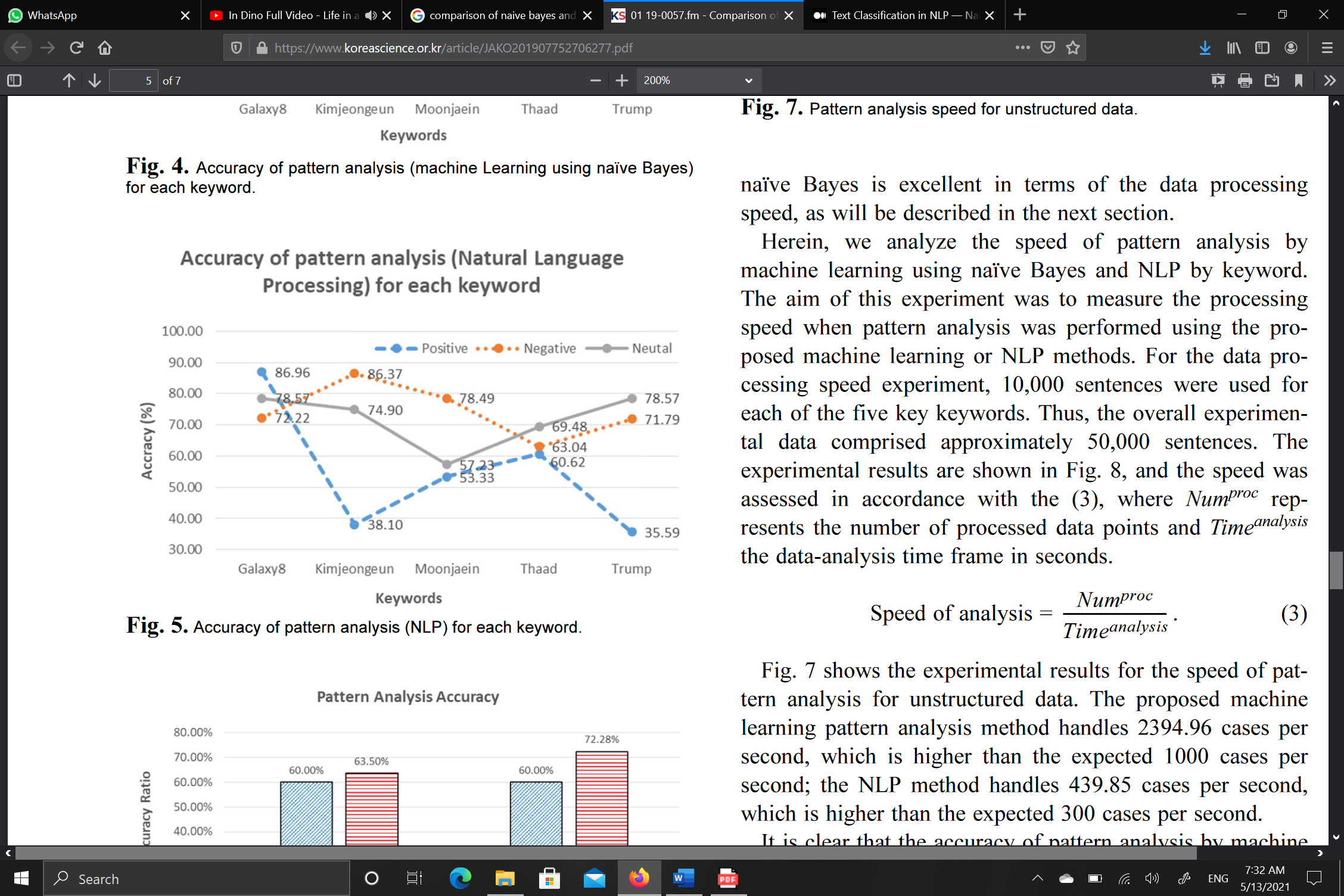
Comparative Study helps to analyze the best the optimal solution for a problem and hence, this chapter enlightens the comparative analysis of Naïve Bayes Algorithm and Natural Language Processing to highlight the optimal algorithm.

**5.1 Comparison of Two Algorithms**

Fig. 4 shows the accuracy of pattern analysis by machine learning for each keyword, while Fig. 5 shows that by NLP. In the pattern analysis method by machine learning using the Naïve Bayes algorithm, the accuracy of the “neutral” opinion was the highest for each keyword, and the pattern analysis method by NLP exhibited the highest accuracy for the “negative” opinion for each keyword.



*Figure 8: Accuracy of pattern analysis (Naïve Bayes)*



*Figure 9: Accuracy of pattern analysis (NLP)*

As shown, the accuracy in the pattern analysis test of unstructured data by machine learning using Naïve Bayes is 63.50%, which is higher than the expected value of 60%. Meanwhile, that by NLP is 72.28%, which is higher than the expected value of 60%.

Therefore, it is clear that the accuracy of pattern analysis by machine learning using naïve Bayes (63.50%) is lower than that by NLP (72.28%).

Table for Comparative Study

|  |  |
| --- | --- |
| **Naïve Bayes Algorithm** | **Natural Language Processing** |
| Naïve Bayes is a probabilistic machine learning algorithm based on the **Bayes Theorem,** used in a wide variety of classification tasks. | Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. |
| Naïve Bayes Algorithm used world cloud to train itself. | NLP uses real time twitter data from Twitter API to train and execute. |
| Naïve Bayes provides an accuracy of 63.50 percent. | NLP provides an accuracy of 72.28 percent. |
| Naïve Bayes Algorithm has less accuracy as compared to NLP. | NLP has better accuracy as compared to Naïve Bayes Algorithm. |
| Naïve Bayes Algorithm bas better efficiency as compared to NLP. | NLP has less efficiency as compared to Naïve Bayes Algorithm. |

# Chapter 6

**Conclusion**

**Chapter 6: Conclusion**

This chapter aggregates conclusion and future scope. The conclusion emphasis on the collective summary of the project and lays the idea about how the project is planned and what is comprises of. The future scope highlights the scope of development in the proposed project with time.

**6.1 Conclusion**

Herein, we analyze the speed of pattern analysis by machine learning using naïve Bayes and NLP by keyword. The aim of this experiment was to measure the processing speed when pattern analysis was performed using the pro-posed machine learning or NLP methods. For the data processing speed experiment, 10,000 sentences were used foreach of the five key keywords. Thus, the overall experimental data comprised approximately 50,000 sentences. It is clear that the accuracy of pattern analysis by machine learning using naïve Bayes is lower than that by NLP; how-ever, it is superior to pattern analysis by NLP in terms of data processing speed. Additionally, the machine learning method using naïve Bayes is less accurate than NLP, but the processing speed is significantly better.

**6.2 Future Scope**

The Existing Database is not able to process the big amount of data within specified amount of time. Also, this type of database is limited for processing of structured data and has a limitation when dealing with a large amount of data. So, the use of Big Data technologies like Hadoop can be used to achieve better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier.

## References

[1] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.

[2] Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019). Sentiment Analysis with NLP on Twitter Data. 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). doi:10.1109/ic4me247184.2019.9036670