

# Sentiment Analysis with NLP on Twitter Data

East West University



Md.Rakibul Hasan

December 2018



# Declaration

I hereby declare that this research project report is an original piece of work carried out by me, under the guidance and supervision of Dr. Mohammad Arifuzzaman. This report is the requirement for the successive completion of BSc in Electronic and Telecommunication Engineering under the department of Electronics and Communications Engineering.

I state that the report along with its literature that has been demonstrated in this report papers, is our own work with the masterly guidance and fruitful assistance of our supervisor for the finalization of our report successfully.

---

Signature of Student:

Md. Rakibul Hasan

ID:2015-2-55-031

Department of Electronics and Communications Engineering

East West University

Dhaka, Bangladesh.

---

Signature of Supervisor:

Dr.Md.Ezharul Islam

Associate Professor

Department of Computer Science and Engineering

Jahangirnagar University

Savar, Dhaka-1342, Bangladesh.

---

Signature of Co-Supervisor:

Dr. Mohammad Arifuzzaman

Assistant Professor

Department of Electronics and Communications Engineering

East West University

Dhaka, Bangladesh.

# Acknowledgement

I would like to express our gratitude and appreciation to all those who gave me the possibility to complete this research work. A special thanks to my supervisor Dr. Mohammad Arifuzzaman, whose help, suggestions and encouragements helped me to take my thesis especially on Data mining, I have been craving to work on it for so long. He supported me by showing different methods of information collection while doing this work. He always helped me when required and he gave required direction towards completion of this work.

I also want to thank all faculty members and staffs of the Department of Electronics and Communication Engineering of East West University for their full cooperation and support during the period of the report completion, from the beginning till the end.

# Contents

<b>1</b>	<b>Introduction</b>	<b>9</b>
1.1	Introduction . . . . .	9
1.2	Twitter . . . . .	10
1.3	Sentiment Analysis . . . . .	11
<b>2</b>	<b>Related Work</b>	<b>13</b>
2.1	Related Work . . . . .	13
2.1.1	Construction Of Word List . . . . .	13
2.1.2	Lexicon-based and Learning-based Methods . . . . .	14
2.1.3	Naive Bayes Algorithm . . . . .	15
2.1.4	Machine Learning Approach . . . . .	15
<b>3</b>	<b>Proposed Work</b>	<b>17</b>
3.1	Natural Language Processing . . . . .	18
3.2	Introduction to Numpy . . . . .	18
3.3	Introduction to Pandas . . . . .	18
3.4	Tokenization . . . . .	19
3.5	Stemming . . . . .	19
3.6	Lemmatization . . . . .	19
3.7	Stop Word . . . . .	20
3.8	Parts Of Speech Predicting . . . . .	20
3.9	Named Entity Recognition . . . . .	21
3.10	Text Modeling . . . . .	22
3.10.1	Bag Of Words . . . . .	22
3.10.2	TF-IDF Model . . . . .	23
3.10.3	Training and Testing Classifier . . . . .	26
<b>4</b>	<b>Twitter Sentiment Analysis</b>	<b>28</b>
4.1	Setting up Twitter Application . . . . .	28
4.2	Twitter Credentials . . . . .	29
4.3	Streaming Connection . . . . .	30

4.4	Fetching Real Time Tweet and Creating Data Frame . . . . .	31
4.5	Setting up the Classifier . . . . .	33
4.6	Preprocessing the Tweets . . . . .	39
4.7	Sentiment Analysis . . . . .	40
4.8	Plotting The Result and Accuracy of our Model . . . . .	42
<b>5</b>	<b>Comparison</b>	<b>45</b>
5.1	Comparison . . . . .	45
<b>6</b>	<b>Conclusion</b>	<b>52</b>
6.1	Conclusion . . . . .	52
6.2	Future Work . . . . .	52
<b>7</b>	<b>Reference</b>	<b>53</b>

# List of Figures

- 3.1 Architectural overview of proposed work . . . . . 17
- 3.2 Tweet . . . . . 26
- 4.1 Twitter API Account . . . . . 28

# Abstract

Every social networking site are generating a huge amount of data in every second. A lots of innovations are coming to the industry to develop new ways of communications between consumer and start-up business opportunity. Sentiment Analysis is the recent research area in data mining. In our work ,we implement the algorithm namely Natural Language Processing(NLP) on social networking data by sentiment analysis. We take twitter as a social networking site and for data analysis takes two product iPhone vs Samsung. Also the comparison between the python built-in library "TextBlob" and Natural Language Processing. The comparison is showing by evaluating the polarity of the tweet as positive and negative towards a particular device. Consumer can be informed their choice according to the general sentiment expression from the other Twitter users on various product.



# Chapter 1

## Introduction

### 1.1 Introduction

Now-a-days, internet services are generating a large amount of data which is increased significantly day by day. Most of the data generated from the social networking site. These social networking site are being used for microblogging where microblogging has become a tremendous tool among the Internet users for communication. Trillions of the user can share trillions of the message as their opinion in different aspects of their day-to-day life in popular social networking website such as Twitter, Tumblr, Facebook, Instagram, Snapchat, Whatsapp, LinkedIn, Youtube. People who post their messages those are writing about their life, share opinions on various topic and discuss about current incident. Now-a-days every big and small company are joining the social networking site to share their product and try to know the reviews of the products from the consumer. They can use those reviews to research public opinion about their product and their company. Consumers can also use sentiment analysis for researching about product or services before consumers are making a purchase.

According to research [? ], Facebook, which is the most famous social network where users like 4 million post in every minute and 250 million post in every hour. Another social networking site Instagram generates 1.7 billion likes from user in a single minute. Twitter is the second biggest social networking platform after Facebook which generates 347,222 tweets every each minute and 21 million tweets per hour. On the other hand, Snapchat users share 284,722 snaps every minute.

Because of a huge amount of data generating from the twitter, it has become very popular and has grown rapidly as a microblogging social networking website. It also creates an opportunity for data mining and sentiment analyses based on users tweet. Since sentiment analyses are part of the data mining that can observe public educe about various topic and product. It may analysis what people think about

their business products and quality of the products, brands, strategies of pricing and is it trendy?.

We are choosing Twitter for sentiment analysis because of it offers the opportunities for the tenderness of enunciated disposition. Twitter is limited to 140 characters of text that's why users can explain their brief ideas via a short message. The sentiment analysis is useful in several ways and contains many branches of computer science such as Natural language Processing, text mining, Information theory and coding, machine learning. The main aim is to identify the sentiment of the tweet by defining positive and negative polarity where tweets are collected by using Twitter streaming API from twitter. From the sentiment, we can show the comparison between two most popular mobile phone device "iPhone" and "Samsung"

## 1.2 Twitter

Since twitter is the second biggest social networking platform where people can share his/her opinion via a short message within 140 character. People can communicate with other people all over the world with many short message in twitter is called tweets. According to twitter website[2] maximum 140 characters of text is the perfect length for sending status from users. Furthermore, people's names are reserved by 20 more characters. To upload tweets and follow other member to know the update about the latest news users must be signed up into service and register for free account. Users can sign up for free account to go the official website of twitter which is <https://twitter.com/>.

Now-a-days social networking websites are not limited only to the website access but also users who can involve and interlude with services those are coming through smart user. According to Twitter usages survey, there are 80% of active users are using Twitter on mobile. The present CEO of Twitter Jack Dorsey try to collaborate inquisitive, productive and creative people all over the world.

Twitter works with several section in their community by building up an open platform which is maintained by a skilled executive team. Business, Developers, help and marketing these are several section where people can advertise their product, company can analyses about their product, consumer may helped by their decision making to purchase a product. In every minute, user's generated data is creating many opportunity for marketing and advertisement. Many companies are using Twitter to keep people's opinions on Twitter about their offers and deals. In textual context of user's tweets has built up a relationship with two meta-data that are "entities" and "places". The entities are provided by the user. User's can mention other users in own tweets by including @ sign that followed by their username is called the way of representation of entities. It also contains hastags(#)

and URLs(<https://>).

An important part of the twitter is Terminology which teaches users about features,access and functionality of the service and also teaches how to use it in an efficient way.Basically @ sign is used to call someone username in tweet or send a message to that user.For an example Barack Obama is @BarackObama[3].Another popular symbol which is used on Twitter is called hashtag(#).It may help users to categorized messages with the # sign which is followed by the related keyword based on its context.Hashtags can be placed at anyplace in the tweet and by using Twitter search it enables better search.

Twitter also supports sending private messages among the users to ensure privacy and security which service is using most of the social networking platform such as Facebook,Viber,WhatsApp.In Twitter this messaging services added extra value of microblogging.

”♡” This Button shows the positive reaction towards the tweets.There are several button to comment,Retweet and to send direct message.People can see other users tweet in their personal Twitter accounts is known as ”following.Those people who receive your Twitter update is called ”follower”.Geotag is used in order to know about the location of the creator during posting tweet.In every posted tweet comment option comes whereas follower can give their reaction and response favour or against that tweet.Retweet option is also a part of every message.By retweet creator can share the tweet with followers.

### 1.3 Sentiment Analysis

Sentiment analysis is known as opinion mining,opinion extraction,subjectivity analysis or emotion AI.Sentiment analysis is a existent research area in data mining.It is the stem of natural language processing, text analysis, computational linguistics, bio-metrics,machine learning methods.Sentiment analysis is an important source which is applied to the voice of customer materials.Sentiment analysis can be extracted,evaluated and identified from the reviews and survey responses, online and social media, and health-care materials.Basically it deals with the user grasp about individual fact.

peoples are sharing their ideas,thinking,opinion with a short message via different blog and social networking platform that brings new scope for developing creative customer service solutions.Sentiment explains analyses of data which is extracted by different technologies and data mining techniques.Now-a-days most of the people all over the world are connected with social network platform such as Facebook,twitter,LinkedIn,reddit where users can express their opinion on particular product,service,brand,political,sports etc.

Insights from the sentiment analyses the company have got idea about the

product how positive or negative are people, political parties can be known how much people support their work, Social organizations and NGOs companies can know people's opinion by questioning. These sentiment analysis is evaluated by classifying the polarity of a text at a document. The classifier is evaluating a document by positive, negative and neutral polarity or showing emotional states such as Happy, sad, angry.

# Chapter 2

## Related Work

### 2.1 Related Work

In this chapter, we are shortly introduced with some data mining or sentiment analysis of Twitter data techniques. The advantage of web technology and its growth a huge bulk of data appear and generated in the web from internet users. Social networking sites have become a platform for online learning, sharing opinion, thinking, ideas where people can share and express their views about affairs and making discussion with various communities across the world. There are several techniques have already introduced to sentiment analysis of twitter data which helps to analyze the information in the tweet. At end of the every techniques sentiment is expressed by showing positive and negative polarity of tweet.

#### 2.1.1 Construction Of Word List

According to this research paper[4], This is the simplest sentiment analysis techniques where compares the words of a posting contrary to a labeled word list where every word has scored for valance. "Sentiment lexicon" and "affective word lists" helps to score for valance of each word. ANEW which is "Affective Norms for English Words" that is developed before advent of opinion mining and microblogging.

Finn Årup Nielsen proposed a new word list that gives better performance than ANEW. According to SentiStrength algorithm, there are two approaches to sentiment analysis. One is the label text data uses supervised machine learning to train & classify the polarity of new texts. Another technique creates a sentiment lexicon and scores that explain how the words and phrases of the text is matched with the lexicon. In SentiStrength software, there are many labeled word lists such as ANEW, General Inquirer, OpinionFinder, SentiWordNet and WordNet-Affect.

Finn Årup Nielsen constructed a new word list with sentiment strength and the subordination of internet slang & obscene words. This method has used for Twitter

data sentiment analysis. To create labeled language data, researchers are using AMT (Amazon Mechanical Turk). AFINN-96 is a word list which has distributed on the web with 1468 different word with a few phrases. SentiStrength uses a scoring range to score the labeled data which is -5 to +5 where (-5) is strong negative and (+5) is strong positive. To discover relevant word Finn Årup Nielsen used Microsoft web n-gram. Sentiment Strength is obtained from SentiStrength by using its web services and converting the positive and negative sentiment by selecting a large numerical value. If the the magnitude of the positive and negative sentiments are equal that would be clustered as a zeroing which is called neutral.

Each tweet has been rated 10 times and each rating has a sentiment strength with an integer between 1 to 9 where 1 is assuming as negative and 9 is positive. Finn Årup Nielsen identified 15,768 word from 1000 tweets where 4095 words are unique. From 4095 words only 422 word hits his word list which contains 2477 and ANEW hits only 398 word of its 1034 words. General Inquirer hits only 358 word of its 3392 words and OpinionFinder hits only 562 words of its 6442 words.

$$\text{CombinedSentimentStrength} = \frac{\sum \text{Valences of the word}}{\text{Number of Word}}$$

So, the newer method is implemented in SentStrength which is using a range of techniques, detection of negation and handling of emoticons & spelling variations.

## 2.1.2 Lexicon-based and Learning-based Methods

In this paper [5], a new entity level sentiment analysis method is proposed for Twitter data where sentiment analysis on entities those are products, organizations and people. They claimed that their method could give high precision and low recall. The proposed method is combining lexicon-based and learning based method.

By using lexicon-based approach in any document or any sentence, the sentence polarity can be determined whether it is positive or negative via some function of opinion word which is called opinion lexicon. Opinion lexicon is a dictionary which contains opinion word to determine & identify the sentimental prediction such as positive, negative and neutral. In their proposed technique, at first they are crawling tweets from Twitter. Then preprocessing the tweets by cleaning noise data such as "RT", external links, user names before sentiment analysis. Then detect the type of sentence those are Declarative, Imperative or Interrogative sentence. The semantic orientation in any sentence or document explain that the infer of opinion or sentence is either positive, negative or neutral. Another way the semantic orientation is a determination of subjectivity and opinion in any document. In opinion rule section state that implication within expression on the left and implied opinion is on the right.

At learning-based method, they train a sentiment classifier to give sentiment polarity and also classifies new tweets those are opinionated. The positive and negative tweets are using as training tweet. Then empirical evaluations determine the comparison between ME, FBS, AFBS, LLS, LMS. At the end of this method, the analysis on sentiment according to work on Twitter data which improves the re-call, F-Score and output forms.

### 2.1.3 Naive Bayes Algorithm

Naive Bayes algorithm is another approach to analyze the sentiment on Twitter data. Naive Bayes algorithm build a Naive Bayes classifier to train and test the ingredient & also determine the sentiment polarity.

After fetching tweets from Twitter API the data should be preprocessed then Chi-squared test used to select feature by finding score words. Then sentiment scoring module is using Naive Bayes classifier to classify tweet as positive or negative[5]. The classifier is followed Bayes theorem,

$$P(c|p) = \frac{P(p|c) * P(c)}{P(p)}$$

where c is class & p is predictor,  $P(p|c)$  is the probability of predictor.  $P(c|p)$  is the probability of posterior where c in conditioned on p that means class holds true given the value of p.

The accuracy of the classifier is calculated by

$$Accuracy = \frac{\sum Positive + \sum Negative}{Total\ number\ of\ words}$$

Where Positive is number of tweets identified as positive & Negative is number of tweets identified as negative.

### 2.1.4 Machine Learning Approach

Machine learning based technique uses some features of classification technique to train the classifier to classify the sentiment perceive unigrams or bigrams. There are two types of machine learning techniques which is unsupervised learning and supervised learning technique.

The unsupervised learning approach doesn't consist of a category & it doesn't provide correct output at all according to rely clustering.

The supervised learning approach is based on label datasets which are trained to provide meaningful outputs. To supervise the learning approach, apply Naive Bayes

algorithm, maximum entropy & support vector machine which helps to achieve great success in sentiment analysis. In machine learning technique, two types of datasets are needed:

1. Training set
2. Testing set

Machine learning is starting with collecting train dataset and apply a classifier to train the dataset. When supervised the sentiment classifier then it selects feature to decision make. Term frequency & their frequency, Negations, Part of Speech information and Opinion words & phrases features are using in sentiment classification.



# Chapter 3

## Proposed Work

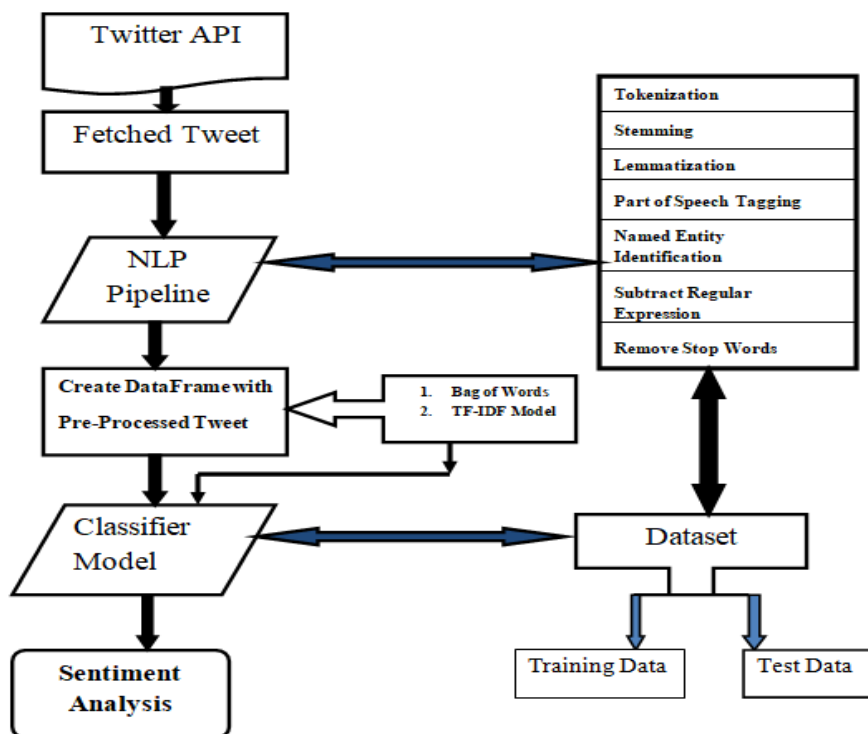


Figure 3.1: Architectural overview of proposed work

## 3.1 Natural Language Processing

From the starting of computer uses, programmers have been trying to write program that can understand language such as English. Basically computer doesn't understand the language in the way that generally humans do. But now-a-days they can do a lot. The process of reading and understand the language is very complex for a computer. Natural Language Processing (NLP) has done it in efficient way where it also deals with text analysis, data mining, spell correction, subtract unwanted symbol & word, create spam classifier and machine translation.

So the Natural Language Processing can be taught our computer to understand as language used by humans. Twitter sentiment analysis can be implemented by using data mining technique with Natural language processing. To create solution with NLP needs some processing tools to understand the language. Natural Language Tool kit (nlTK) is open source library which is developed in python to apply NLP techniques. These tools are containing necessary tools to text process and sentiment analysis.

## 3.2 Introduction to Numpy

NumPy is a linear algebra library or fundamental package for scientific computing in python. Almost all of the libraries at pydata ecosystem in python depend on NumPy. If you have Anaconda, you can install NumPy by going to your command prompt by using "pip install numpy"

It contains a powerful N-dimensional array object, many broadcasting functions, using tools for integrating C/C++ and Fortran code. It can be used for linear algebra, Fourier transform, and random number generating. At NumPy array has two component i.e Vector & Matrix where vector is 1-D and matrix is 2-D.

## 3.3 Introduction to Pandas

Pandas is a powerful data analysis python toolkit. This is a software which is written for python to data manipulation and data analysis. It provides fast, flexible and expressive data structures. Pandas is well suited for tabular data, arbitrary matrix data & observational or statistical data sets.

Pandas does well to handle missing data, automatic & explicit data alignment, column insertion & deletion, easy to convert python & NumPy data structure into data frame object.

## 3.4 Tokenization

At Natural Language processing pipeline, the first step is sentence Tokenization & second step is word Tokenization. In the first step, the document will split with separate sentence. Then it will be easier to a programmer to write a program that can be understand a single sentence then understand the whole sentence. In the second step, the sentence break into separate words which is called "token" & the process is called Tokenization.

Input: Each new language presents with some new issues

Output: "Each", "new", "language", "presents", "some", "with", "new", "issues"

This is an easy way which is done with English. It will split apart word when it gets space between words.

## 3.5 Stemming

Stemming is an important concept of Natural language Processing that comes when extracting some features out of mere sentence or corpus of a lot of sentence. The process of stemming in any sentence the inflected word is reducing to their base or root form.

Suppose there are some word which occurs inside different sentence in a document such as "Intelligently", "Intelligence", "Intelligent" the base form of those word is "Intelligen". Another way "goes", "going", "gone" is converted into "go". This is called stemming. So, we are not duplicating different words when we have the same word in different forms. We are not taking those word as different words rather than we are taking those word as same word. It takes less time to analysis a sentence or a document. Stemming is used where meaning isn't important such as spam detection in a content. We can import library from nltk by using

```
import nltk
from nltk.stem import PorterStemmer
```

## 3.6 Lemmatization

In previous section which is talking about stemming we can see a base or root form is "Intelligen" that doesn't make any meaning. So, we can figure out that intermediate representation of the word may not have any meaning. When working in a computer with text, it helps to know the root form of the word which explain that the different sentence are explaining the same concept.

So the lemmatization is same as stemming but intermediate representation of word in base form which has contained some meaning.If you are doing some kind of analysis where meaning is important then you can use lemmatization.It takes more time to analysis and this process is using where meaning of a word is important such as question answering applications.We can import library from nltk by using

```
import nltk
from nltk.stem import WordNetLemmatizer
```

### 3.7 Stop Word

Many different words that appear a lot of times in different sentences such as "and", "The", "a", "be", "to" etc these are the common word whether those words have pretty much no meaning when extracting any features from any sentence and creating a lots of noise since they are appearing more frequently than other words in a document.These common words are called stop words.

So stop words have no meaning because they are not able to express any special meaning based on some specific context.so when we are doing sentiment analysis these words have no impact on sentiment whether the sentiment is positive or negative.This is the reason in most of the cases we really want to remove these different stop words to get better performance.You can download stop word package by using

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

Basically stop words are identified from the list of known stop words & there is no standard list of stop words which is suitable for all application.It can vary upon depending on your application.

### 3.8 Parts Of Speech Predicting

In this pipeline,program is looking each tokenize word and gives its part of speech.Knowing the part of speech of each word in the document will help to understand what the sentence is talking about.The model of par of speech was trained by millions of English sentence with its each word's POS so that it can learn the behaviour of that word.After Tokenization each tokenize word can show their POS by using

*nltk.pos – tag(words)*

The meaning of Part of Speech in NLTK:

- CC means Coordinating conjunction
- CD means Cardinal number
- DT means Determiner
- EX means Existential There
- JJ means Adjective
- JJR means Adjective, comparative
- JJS means Adjective, superlative
- NN means Noun, singular or mass
- NNS means Noun, plural
- NNP means Proper noun, singular
- NNPS means Proper noun, plural
- POS means Possessive ending
- RB means Adverb
- RBR means Adverb, comparative
- RBS means Adverb, superlative

### **3.9 Named Entity Recognition**

By using NLP, we can extract a list of real-world from a document. Named Entity Recognition is used to detect and label the noun word with the real world concepts. This is using in the context to find out how a word appears in the sentence and it gives a statistical model to guess which type of noun appears in a word.

A good Named Entity Recognition can identify the "Brooklyn Decker" as a name and "Brooklyn" as a name of place. NER system can tag some kind of object those are given below

- People's Name identify
- Company's Name identify

- Product's Name identify
- Geographical Location for both physical and political
- Event's name
- Display Date & Time

## 3.10 Text Modeling

Natural Language Processing is the subarea of computer science and artificial intelligence which refers the normal language that we use to communicate to converse language which will understand computer such as computer code. In natural language processing, Text modeling which is also called statistical model to use for discovering the abstract from the document. Basically Text modeling is using as a text mining tool to find out semantic structure from a text body. We are using some text model to evaluating sentiment analysis by building a classifier. We are testing by two model to train & testing the classifier which gives us more accuracy.

- Bag of Words
- TF-IDF Model

### 3.10.1 Bag Of Words

In Natural Language Processing, Bag of Words is a representation of a sentence or a document as a multiset of its words without following any grammatical rule or words order and keeping those words in multiset as a multiplicity. Where multiplicity is a member of multiset that counts how many times the words appears in multiset & represent its as a binary number. If the word presents in the sentence from the whole corpus then it will give one otherwise it gives. And it will make matrix.

Bag of Words(BoW) is an algorithm which is calculating the number of times a words appear in a document. Those word counts give us a infer to compare between the document and gauge how much similarity presents in those document. The Bag of word can be used to document classification & text modeling. We can feed this model to any machine learning algorithm to analyze any sentence or any document. We can follow this process to classify a huge corpus of data. Example:

Here is given two sentence:

- Today is going to rain.
- I am not going today

So we need to convert all these different sentences into an intermediate presentation on a model and then we can feed that model in any algorithm to do any kind of analysis. In above, we have already removed the different punctuation marks which isn't necessary. Now we will convert all sentence into lower case

- today is going to rain.
- i am not going today.

Now Create a Bag of words from the document of different sentence:

Bag of Words							
Words/Documents	going	is	to	today	i	am	rain
sentence1	1	1	1	1	0	0	1
sentence2	1	0	0	1	1	1	0

So, in the columns there are all the frequent words and in the rows there are all the different documents and ones and zeros corresponding to whether the word appears in the document or the word does not appear in the document.

### 3.10.2 TF-IDF Model

TF-IDF means Term Frequency-Inverse Document Frequency. Before discuss about TF-IDF, we have to know what's the problem that appears on the BoW model which is explained on the previous section. We will build a model which is more efficient compared to the Bag of Words models. We have already see, there are different words in the different documents and the word appears in a document we will put one over model and if word does not appear in the document then put a zero. So this is the logic of the whole bag of words. Let's find out the problem

- It gives all words have the same importance.
- It doesn't preserve any semantic information.

So, when we feed this model into the machine learning algorithm then the machine will be confused that all words have the same impact on the document. Suppose, The girl looks pretty then the word "pretty" is important in this sentence. So, we can improve our Bag of Words model by using "TF-IDF" Model.

Term Frequency-Inverse Document Frequency is numerical statistics that implications how important a word in a collection of document. It measure relevance not frequency. Where TF (Term Frequency) is a frequency of a particular word in a particular document.

$$\text{TermFrequency, } TF = \frac{\text{NumberofOccurencesofWordInDocument}}{\text{NumberofWordsInThatDocument}}$$

And IDF(Inverse Documnt Frequency) isn't calculated per document rather than it's calculated for the whole document & it is the inverse document frequency of a word in the whole documents.It gives a single idea each of the word.

$$\text{InverseDocumentFrequency, } IDF = \ln \frac{\text{NumberOfDocuments}}{\text{Numberofdocumentscontainingword}}$$

So,

$$TF - IDF = TF * \ln \frac{\text{NumberOfDocuments}}{\text{NumberOfDocumentsContainingWord}}$$

It also represents by

$$TF - IDF = TF(\text{Document}, \text{Word}) * IDF(\text{Word})$$

Example:Taking two sentence

- today is going to rain.
- i am not going today.

The frequency of most frequent word:

Frequency List	
Word	Frequency
today	2
is	1
going	2
to	1
rain	1
i	1
am	1

Now the Term Frequency for each of the different words in the histogram of frequency:



Term Frequency		
Word	Sentence1	Sentence2
today	0.2	0.2
is	0.2	0
going	0.2	0.2
to	0.2	0
rain	0.2	0
i	0	0.2
am	0	0.2

According to formula of Inverse Document Frequency the IDF values are:

Inverse Document Frequency	
Word	IDF Values
today	0
is	0.693
going	0
to	0.693
rain	0.693
i	0.693
am	0.693

Now we have the values of TF & IDF model. Then calculating the final value of TF-IDF model which is given below

TF-IDF Values							
Words/Documents	going	is	to	today	i	am	rain
sentence1	0	0.14	0	0.14	0	0	0.14
sentence2	0	0	0	0	0.14	0.14	0

The bag of words contains only zeros & ones. But this model contains zeroes because of the idea of values but it contains these fractional values also such as 0.14. So when we are doing it on a huge corpus of data we have this bunch of decimal values and if we look closely at these different values then we will be able to find out that most of the important words in the whole document have a higher value and the higher fractional value of a document. So the word "rain" is a very important word in our two document corpus. So it has a very high value which is 0.14 & so on. So you can see in the word "going" is a very common word right because it appeared in all the documents. So it got the lowest TF-IDF value which is zero. So similarly in this way in the TF-IDF model can give more importance to some specific word and that is the reason why this model is extensively used in case of text classification are in our opinion mining and many other applications.

### 3.10.3 Training and Testing Classifier

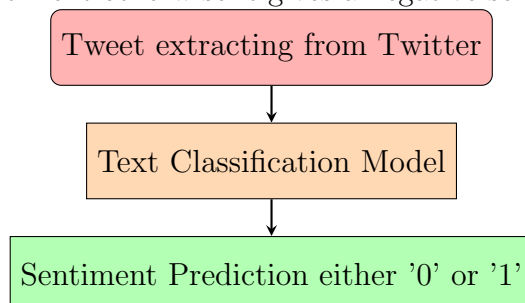
In Twitter a lots of user posted tweet those are messy,unstructured & sometimes that tweet doesn't contain nonsensical meaning.The challenging part is extracting data from messy,unstructured text that doesn't follow any grammatical rule and analyzing its grammatical structure.

Now look at an user review which is one of the most common type of data you want to parse in the computer.Here is given an public tweet about Samsung Electronics Gadget.

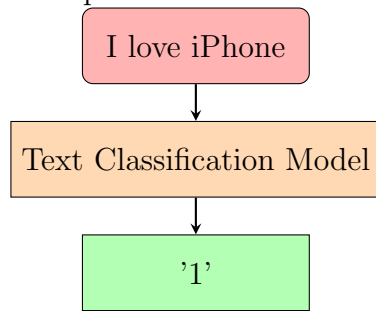


Figure 3.2: Tweet

From the screenshoot,you could automatically understand that it is a positive tweet about Samsung electronics gadget.So how can we write a program for this text that will understand & give a positive feedback or sentiment among this tweet.To get the sentiment of this tweet we set up a linear classifier that takes in word.The text of the tweet will be the input of the classifier.The output will be '0' or '1' where '0' stands for negative tweet & '1' stands for positive tweet.If the tweet is positive then the classifier gives the feedback that the tweet has a positive sentiment otherwise it gives a negative sentiment.We can explain it by a flow chart:



If the classifier can understand the text and reliably predict the correct sentiment that means the classifier somehow understand the overall meaning from extracting tweet. The classifier isn't similar to the human intelligence but even if it gives the prediction which is approximately or better than human intelligence then it doesn't matter. To get better accuracy from our classifier we need to train up our text classification model. When the model is perfectly trained then we can use this model to make prediction for next text document.



In chapter 4, I am discussing about text classifier model broadly to predict sentiment from extracting tweet.

# Chapter 4

## Twitter Sentiment Analysis

### 4.1 Setting up Twitter Application

Access to the Twitter service by using two methods those are Streaming API and REST API. Streaming API provides the access to get service when tweets are fetching as a continuous stream of information. Rest API is using in representational state transfer. Basically we are using Streaming API for collecting data from Twitter to analyze sentiment on Twitter data.

In order to collect data from Twitter, we need to create a Twitter application that will help to obtain access token, consumer key, API secret. To access data from other services without any credential we need to open OAuth that provides the capability. At first, after getting access token to create a new application from Twitter Application Management.

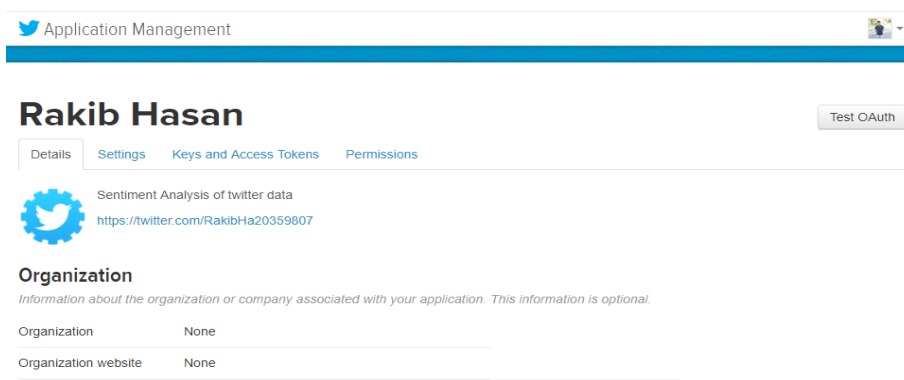
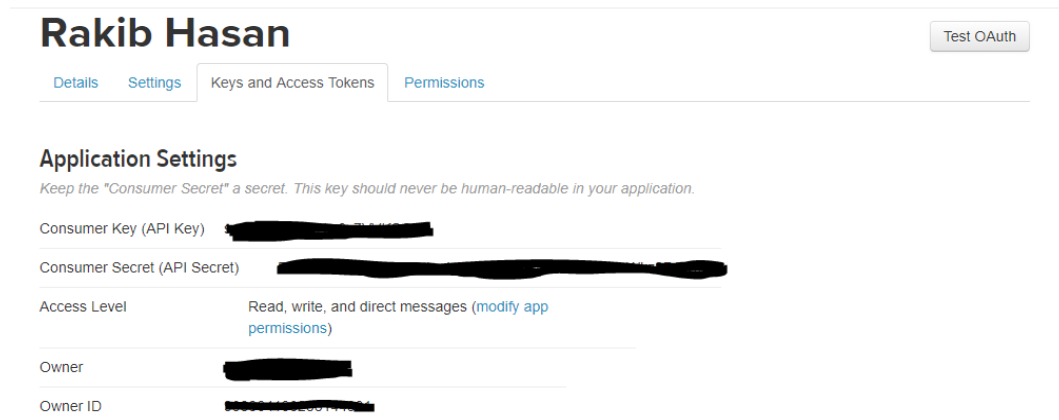


Figure 4.1: Twitter API Account

## 4.2 Twitter Credentials

In Twitter, there are four credentials: Consumer Key (API key), Consumer Secret (API Secret), Access Token, and Access Token Secret [9]. These four credentials provide everything for Twitter data mining to authorize and create an API request on behalf of its owner. API key and Secret provides the access level to write and read Twitter data.



**Rakib Hasan** Test OAuth

[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

### Application Settings

*Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.*

Consumer Key (API Key) [REDACTED]

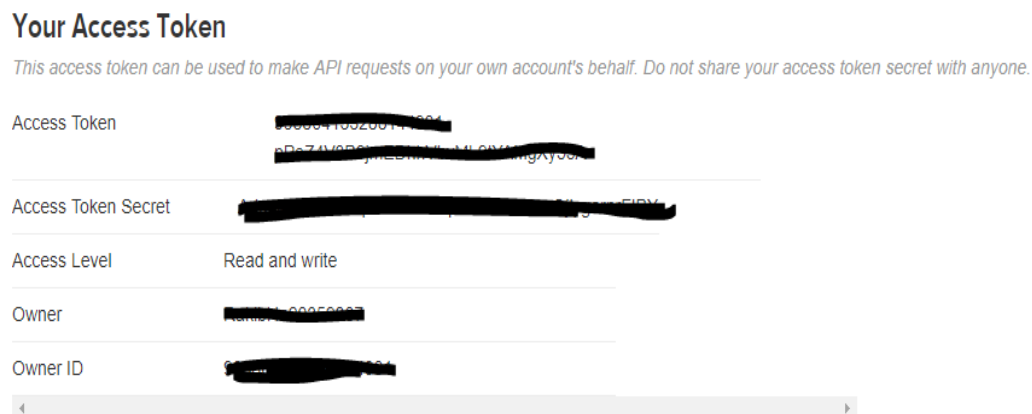
Consumer Secret (API Secret) [REDACTED]

Access Level: Read, write, and direct messages ([modify app permissions](#))

Owner: [REDACTED]

Owner ID: [REDACTED]

Another two credentials, Access Token & Access Token Secret, are used to create an API request to the Twitter service from the owner ID.



### Your Access Token

*This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.*

Access Token: [REDACTED]

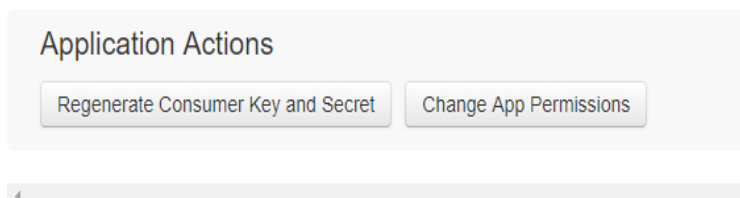
Access Token Secret: [REDACTED]

Access Level: Read and write

Owner: [REDACTED]

Owner ID: [REDACTED]

If you want to regenerate or revoke your Application Actions or Token Actions, you can easily regenerate from the application.



### Application Actions

[Regenerate Consumer Key and Secret](#) [Change App Permissions](#)

## 4.3 Streaming Connection

When credentials are successfully obtained by application management or Streaming API. At first we have to create a connection with Twitter API to collect tweet from Twitter in order to analyze of each tweet & their attributes.

At first we are importing some library to access one module of python code from another module. The import system helps to invoke the import machinery. Basically, the import statement combines of two operation. At first, it searches for the named module then it binds the results of search to a name into a local variable.

```
In [1]: import tweepy
import re
import pickle
import pandas as pd
import numpy as np
```

Here we are importing "Tweepy" to consume Twitter's API. And "import re" for Regular expression which specifies a collection of string matches with the given regular expression. Then "import pickle" is used to serialize python object. "import pandas" uses to handle data and "import NumPy" uses for computing number.

For Plotting and visualization, some another library those are "import display", "matplotlib.pyplot", "seaborn"

```
from IPython.display import display
import matplotlib.pyplot as plt
import seaborn as sns
```

From tweepy import OAuthHandler to create object auth in order to set up authentication and have to be passed Consumer Key & Consumer Secret to OAuthHandler. It will successfully authenticate our set up when set\_access\_token is setting up "Access Token" & "Access Token Secret"

```
# Consume:
CONSUMER_KEY = 'sx611f13gRsOBnw3y7VVKSQT7'
CONSUMER_SECRET = 'DSZiXlmHEsKTcT7JMnahQY2dPC4lQNV21hkorPszWkz8D2Q5oH'

# Access:
ACCESS_TOKEN = '998804133288144901-pPe24V3B9jmEDhhVbvML9tYAMgXy55A'
ACCESS_SECRET = 'Adqyik2TdUaStqYAF4hKzCpRwSiLHWveOjlygarnrFIBY'

# Authentication and access using keys:
auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_TOKEN, ACCESS_SECRET)
```

So, when we have built a Twitter application that by default the access tokens and access secrets are not generated. After generate Access Token and Access Token secret and pass it over `auth.set_access_token(ACCESS_TOKEN, ACCESS_SECRET)` then it can fetch tweets from Twitter.

## 4.4 Fetching Real Time Tweet and Creating Data Frame

For analyzing data we need some sample of tweets. And we created two object those are "args" & another is "api". Now we are going to fetch some sample tweet from the Twitter corresponding to "Iphone". We are going to fetch the top hundred recent tweets about Iphone from Twitter. The process of the collecting and analyzing tweets is one of the most important part of data mining. And collected Tweet from Twitter store in a dataset which is human readable dataset. At first, we declare the list which is called "list\_tweets" and it contains all the 100 tweets about Iphone.

All fetched tweets are printed in console which is given below

```
print('We display the first 10 elements:')
display(data.head(10))
```

```
We display the first 10 elements of the dataframe:
      Tweets
0  I was 40 years old when I found this out. On a...
1  Don't Miss the Best Cyber Week Deals on iPhone...
2  @tolu_tezzy iPhone X
3  Thanks for patronizing Darling ❤️\nLumee Duo fr...
4  Check out MMOBIEEL Home Button for iPhone 6S / ...
5  @babiaannabell 🤔🤔🤔 will someone ever have peac...
6  Did you know meditation has significant benefi...
7  Your mood reall affects how you see the world....
8  @gustdyg i wasnt gonna buy that iphone until u...
9  Do you trust your swing enough to put your iPh...
```

To show the storing tweet in dataframe of "list\_tweets" creating a pandas dataframe

```
# We create a pandas dataframe as follows:
data= pd.DataFrame(data=[tweet for tweet in list_tweets], columns=['Tweets'])
```

Index	Type	Size	Value
0	str	1	I was 40 years old when I found this out. On an iPhone press and hold ...
1	str	1	Don't Miss the Best Cyber Week Deals on iPhone XS, iPhone XR, iPhone X ...
2	str	1	@tolu_tezzy iPhone X
3	str	1	Thanks for patronizing Darling ❤️ Lumee Duo from us
4	str	1	Check out MMOBIEI Home Button for iPhone 6S / 6S Plus (Black/Space Gre ...
5	str	1	@babiaannabell 🙄🙄🙄 will someone ever have peace with an iPhon...
6	str	1	Did you know meditation has significant benefits on your mental?
7	str	1	Your mood reall affects how you see the world. I've been mad at everyt ...
8	str	1	@gustdyg i wasnt gonna buy that iphone until u told me too uwu
9	str	1	Do you trust your swing enough to put your iPhone out in front of you ...
10	str	1	8 BALL POOL BLOG #252("4 DOUBLE POTS IN A ROW")#8BallPool #8ball #game ...

Internal Method for a single tweet:

```
Internal methods of a single tweet object:
['_add_', '_class_', '_contains_', '_delattr_', '_dir_', '_doc_', '_eq_', '_format_',
'_ge_', '_getattr_', '_getitem_', '_getnewargs_', '_gt_', '_hash_', '_init_',
'_init_subclass_', '_iter_', '_le_', '_len_', '_lt_', '_mod_', '_mul_', '_ne_',
'_new_', '_reduce_', '_reduce_ex_', '_repr_', '_rmod_', '_rmul_', '_setattr_',
'_sizeof_', '_str_', '_subclasshook_', 'capitalize', 'casefold', 'center', 'count', 'encode',
'endswith', 'expandtabs', 'find', 'format', 'format_map', 'index', 'isalnum', 'isalpha', 'isdecimal',
'isdigit', 'isidentifier', 'islower', 'isnumeric', 'isprintable', 'isspace', 'istitle', 'isupper',
'join', 'ljust', 'lower', 'lstrip', 'maketrans', 'partition', 'replace', 'rfind', 'rindex', 'rjust',
'rstrip', 'rsplit', 'rstrip', 'split', 'splitlines', 'startswith', 'strip', 'swapcase', 'title',
'translate', 'upper', 'zfill']
```

The internal method represents the meta data which contained in a single tweet. If we want to know the geotag or source of creation & the creation time and date.

We have already known that Twitter is supported only 140 characters to give any opinion or posting any tweet on Twitter. Now we create a dataframe a list of the tweet with its length.

```
# We add relevant data:
data['len'] = np.array([len(tweet) for tweet in list_tweets])

# We extract the mean of lengths:
mean = np.mean(data['len'])
print("The average length of Tweets", mean)
```

The average length of Tweets 115.0



data - DataFrame

Index	Tweets	len
0	I was 40 years old when I found this out. On an iPhone press and hold on the space bar and move your finger to move... <a href="https://t.co/vBRMKGsyIt">https://t.co/vBRMKGsyIt</a>	140
1	Don't Miss the Best Cyber Week Deals on iPhone XS, iPhone XR, iPhone X, and iPhone 8 <a href="https://t.co/Ipclmd7B5a...">https://t.co/Ipclmd7B5a...</a> <a href="https://t.co/Nn5tIysaPV">https://t.co/Nn5tIysaPV</a>	133
2	@tolu_tezzy iPhone X	20
3	Thanks for patronizing Darling ❤️Lumee Duo from us	137
4	Check out MWOBIEL Home Button for iPhone 6S / 6S Plus (Black/Space Grey) ... by MWOBIEL <a href="https://t.co/kthrXw2UWq">https://t.co/kthrXw2UWq</a> via @	117
5	@babiaannabell 🙄🙄🙄 will someone ever have peace with an iPhone? When ...	140
6	Did you know meditation has significant benefits on your mental? For a lot of us It can feel hard to find the time... <a href="https://t.co/hQqt70Zwyp">https://t.co/hQqt70Zwyp</a>	140
7	Your mood reall affects how you see the world. I've been mad at everything lately, fixed my iPhone 2day. I could kiss you	121
8	@gustdyg i wasnt gonna buy that iphone until u told me too uwu I love you too 🥰	81
9	Do you trust your swing enough to put your iPhone out in front of you while you hit balls? What does it take to dev... <a href="https://t.co/nZ109nctFN">https://t.co/nZ109nctFN</a>	140
10	8 BALL POOL BLOG #252("4 DOUBLE POTS IN A ROW")#8BallPool #8ball #gamestagram #iphone #blog #Gaming #miniclip... <a href="https://t.co/grFHDhOWEF">https://t.co/grFHDhOWEF</a>	134
11	@bebizzzy A1: A new laptop or new phone I'm ready to upgrade my 2012 lap t...	139
12	@tolu_tezzy @skizyman iPhone b2	31
13	@JoeSilverman7 -sent from iphone	32
14	@bbceastenders so did anyone else wonder how Sharon read her text when the phone was clearly upside down?... <a href="https://t.co/nEtwDgTFVK">https://t.co/nEtwDgTFVK</a>	130
15	@dipaolamomma Q1 I actually got the new iPhone 8plus! For my birthday last week :) #VZParent #MobileLiving	106
16	Iphone at 7 am:	15
17	Make Sure to Get Your \$29 iPhone Battery Replacements Soon as Apple's Discount Program is Set to End on December 31... <a href="https://t.co/zxKx35oJ50">https://t.co/zxKx35oJ50</a>	140
18	dude i got a new iphone and didn't realize there was no home button how does this shit work	91
19	Anker PowerLine+ II - The Indestructible iPhone Charging Cable <a href="https://t.co/3tRAL1IDWj">https://t.co/3tRAL1IDWj</a>	86
20	Thinking about buying an iPhone 7 or 6s. Had this @oneplus 2 since launch and it's gotten so slow.	98

## 4.5 Setting up the Classifier

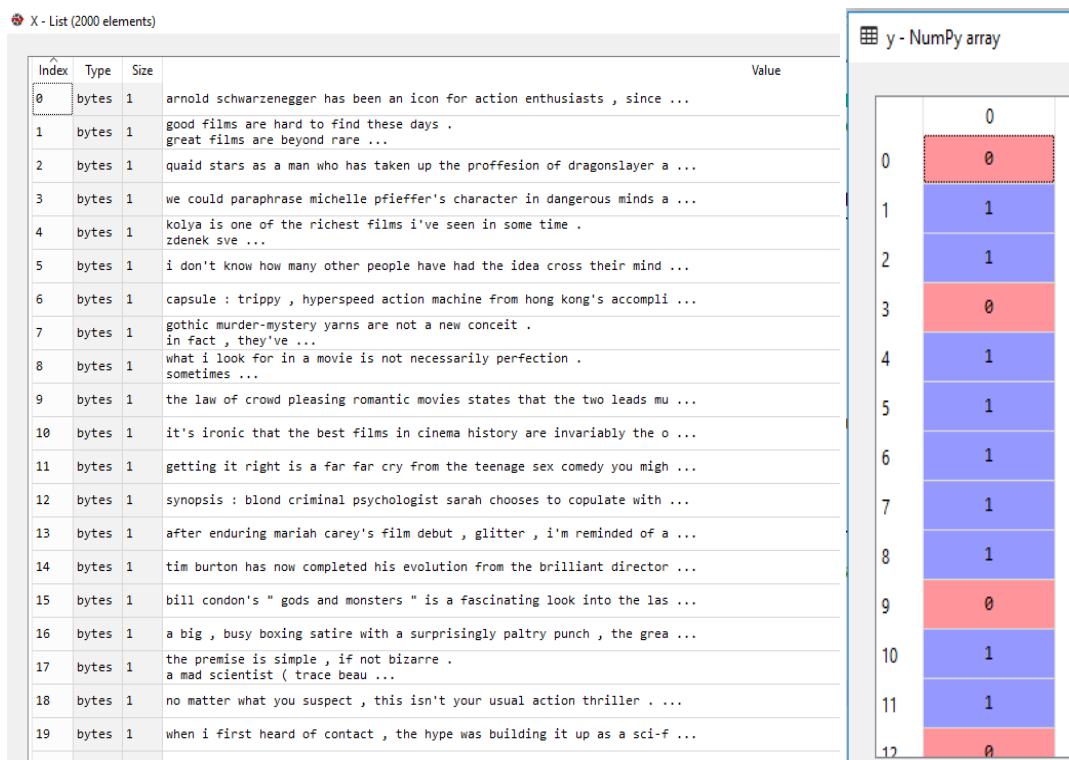
The classifier is a part of machine learning approaches which are analyzed on the Twitter dataset. Tweets were fetching from Twitter & stored in a database which also called dataframe. To predict the sentiment of a text we need to build a classifier. we are going to fetch data from an external source the preprocess that data. After that create an intermediate representation of that data and finally use that data to train machine learning classifier and that classifier would be able to predict whether a given sentence is positive or negative. Now we are going to use our classifier to classify real time tweets from Twitter & this classifier will take the classification is an application of natural language processing by which can create a model and that model is able to classify human language into different classes. To build the classifier we need some data and using Cornell sentiment analysis data set [8]. I have downloaded this datasets. This is a polarity data set which contains about five thousand three hundred and thirty one positive and same number of negative datasets. Now we need some bunch of different libraries to build the classifier.

```

import numpy as np
import nltk
import re
import pickle
from nltk.corpus import stopwords
from sklearn.datasets import load_files
nltk.download('stopwords')

```

Now importing the dataset from the downloaded file which must be in the same folder in python code. This file is going to generate two classes which are X & y. For a neg from the downloaded dataset that it will generate a class zero for a pos from the downloaded dataset it was generated class 1. So when the files are loaded, we need to separate the class and the document and the documents to be in a separate list. Also the corresponding classes to be in another list. Those are X & y where X stands for the list of data set and y stands for the polarity of the each data of the dataset.



Now we are going to persist the data set since we are only doing the text classification on about 2000 dataset. If we want to do the classification on about 50000 dataset and it will take more time. So, we are using less dataset so that the process will be faster. Now store the data set as a pickle file where pickle is using to serialize or de-serialize the python object. Basically, it converts any python object into character stream. At first pickle is serializing the dataset then it writes to a file.

```

#Storing as pickle file:
with open('X.pickle','wb') as f :
    pickle.dump(X,f)

with open('y.pickle','wb') as f :
    pickle.dump(y,f)

#Unpickling the dataset
with open('X.pickle','rb') as f:
    p=pickle.load(f)

with open('y.pickle','rb') as f:
    y=pickle.load(f)

```

Then Preprocess the dataset and create a dataset which named is corpus.This corpus contains all the data in a pre-processed manner.

```

#Creating The corpus
corpus=[]
for i in range(0,len(X)):
    review = re.sub(r"\W", ' ',str(X[i]))
    review =review.lower()
    review=re.sub(r"\s+[a-z0-9]\s+", " ",review)
    review=re.sub(r"^[a-z0-9]\s+", " ",review)
    review=re.sub(r"\s+", " ",review)
    corpus.append(review)

```

Pre-processed dataset:

Index	Type	Size	Value
0	str	1	arnold schwarzenegger has been an icon for action enthusiasts since t ...
1	str	1	good films are hard to find these days ngreat films are beyond rare n ...
2	str	1	quaid stars as man who has taken up the proffesion of dragonslayer af ...
3	str	1	we could paraphrase michelle pfiiffer character in dangerous minds an ...
4	str	1	kolya is one of the richest films ve seen in some time nzdenek sverak ...
5	str	1	don know how many other people have had the idea cross their mind tha ...
6	str	1	capsule trippy hyperspeed action machine from hong kong accomplished ...
7	str	1	gothic murder mystery yarns are not new conceit nin fact they ve been ...
8	str	1	what look for in movie is not necessarily perfection nsometimes movie ...
9	str	1	the law of crowd pleasing romantic movies states that the two leads m ...
10	str	1	it ironic that the best films in cinema history are invariably the or ...

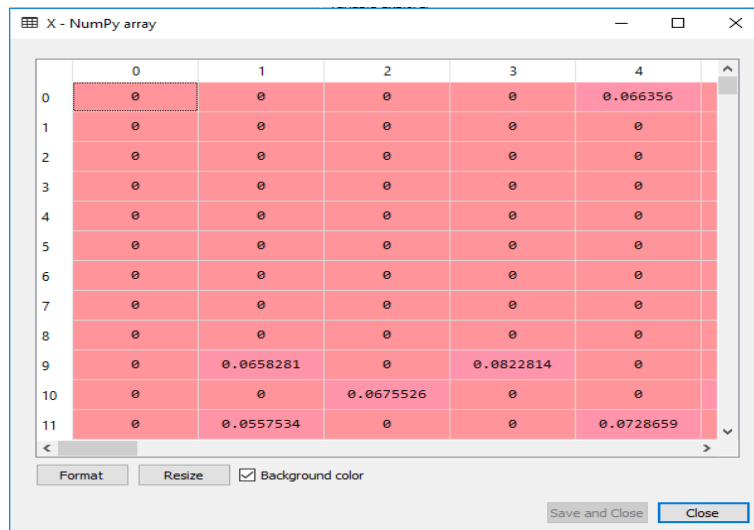
Now create a bag of words model and build TF-IDF model as the improvement model of bag of words and the reason was discussed on previous section

```
#Creating the simple binary bag of model:
from sklearn.feature_extraction.text import CountVectorizer
vectorizer=CountVectorizer(max_features =2000,min_df = 3,max_df =0.6,stop_words
X=vectorizer.fit_transform(corpus).toarray()

#TF-IDF Model
from sklearn.feature_extraction.text import TfidfTransformer
transformer=TfidfTransformer()
X=transformer.fit_transform(X).toarray()
```

tion[3.10.1][3.10.2].

TF-IDF Model:

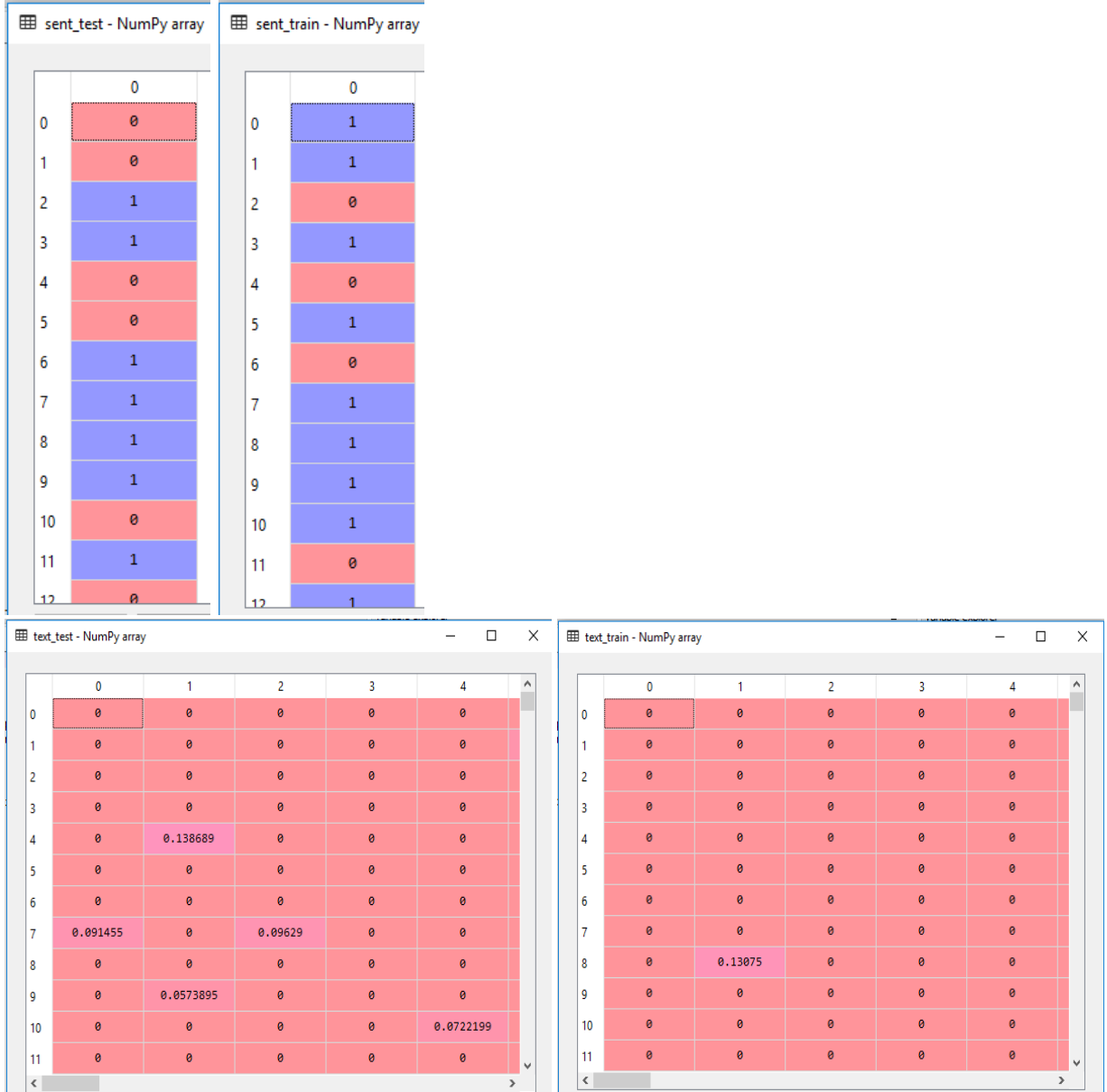


The values of TF-IDF model gives us an idea about which words are important in this document.

Now we are going to split our dataset into a training set and a test dataset. We have 2000 different documents out of these two thousand different documents we want to use some documents for training and rest of them for testing. We will use about 16 hundred different documents to train the model and about 400 documents to test them on a performance. The whole dataset splits for training by using "Sklearn" Python library.

```
#creating training and test dataset
from sklearn.model_selection import train_test_split
text_train,text_test,sent_train,sent_test = train_test_split(X,y,test_size = 0.2,random_state = 0)
```

Now see the data set of training and testing:



Here, the testing text & training text contains about 400 rows and 2000 features so the number of features is obviously constant but there are some varying number in rows. Another important tool is logistic regression which is used to build a classifier to predict the sentiment of a tweet or a document. Basically Logistic Regression has been done some job those are

- It calculates the coefficient values.
- Every new sentence are given a point by calculating coefficient values.
- If the point is greater than 0.5 then the sentence will be predicted as positive sentiment otherwise it evaluates the sentence as negative.

So,we are going to train our model using the text\_train and the sent\_train and use logistic regression algorithm to create the whole classifier.

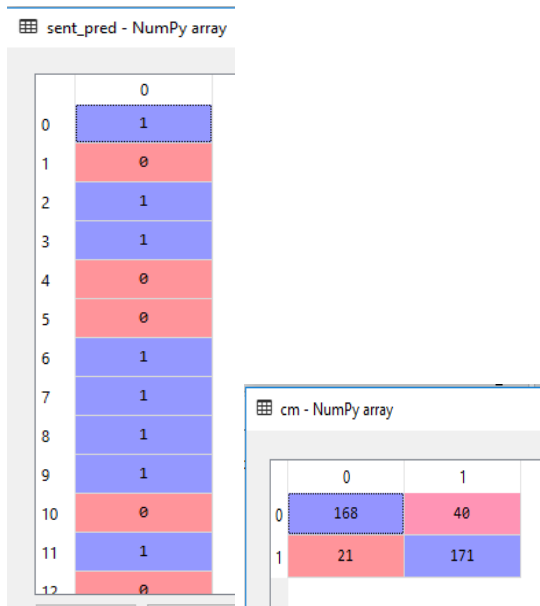
```
#Logistic regression

from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression()
classifier.fit(text_train,sent_train)
```

Now,We have used this logistic regression class and created an object to train our model and evaluate how efficient or accurate our model performance is?To evaluate the accuracy of our model create a confusion matrix according to 400 test data

```
sent_pred = classifier.predict(text_test)

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(sent_test,sent_pred)
```



Where row stands for the predicted value and column stands for the actual value.The confusion matrix evaluates that 168 sentence are negative where our model also predict those sentence as negative and 21 actually negative but our model predict as positive.On the other hand,171 sentence are positive where our

model also predict those sentence as positive and 40 actually positive but our model predict as negative.

Now create pickle for "TF-IDF Model" and "Classifier" those TF-IDF model and classifier using in core code of sentiment analysis.

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer=TfidfVectorizer(max_features =2000,min_df = 3,max_df =0.6,stop_words=stopwords.words('english'))
X=vectorizer.fit_transform(corpus).toarray()

with open ('tfidfmodel.pickle','wb') as f :
    pickle.dump(vectorizer,f)
```

## 4.6 Preprocessing the Tweets

In previous section,we have built an classifier to predict the sentiment.Now we are going to use that TF-IDF Model & Classifier from "tfidfmodel.pickle","classifier.pickle".

```
#Loading TF-IDF model & Classifier

with open ('tfidfmodel.pickle','rb') as f:
    vectorizer =pickle.load(f)
with open ('classifier.pickle','rb') as f:
    clf=pickle.load(f)
```

Now preprocessing the dataframe of "list\_tweets" accroding to Natural Language Processing.The process of NLP was elaborately discussed in the previous section[Chapter:3]

```
for tweet in list_tweets:
    tweet = re.sub(r"https://t.co/[a-zA-Z0-9]*s", "", tweet)
    tweet = re.sub(r"s+https://t.co/[a-zA-Z0-9]*s+", "", tweet)
    tweet = re.sub(r"s+http://t.co/[a-zA-Z0-9]*$", "", tweet)]
    tweet=tweet.lower()
    tweet=re.sub(r"rt", "", tweet)
    tweet=re.sub(r"@[\s]+", ' ', tweet)
    tweet=re.sub(r"that's", "that is", tweet)
    tweet=re.sub(r"there's", "there is", tweet)
    tweet=re.sub(r"what's", "what is", tweet)
    tweet=re.sub(r"it's", "it is", tweet)
    tweet=re.sub(r"who's", "who is", tweet)
    tweet=re.sub(r"i'm", "i am", tweet)
    tweet=re.sub(r"she's", "she is", tweet)
    tweet=re.sub(r"he's", "he is", tweet)
    tweet=re.sub(r"they're", "they are", tweet)
    tweet=re.sub(r"who're", "who are", tweet)
    tweet=re.sub(r"ain't", "am not", tweet)
    tweet=re.sub(r"wouldn't", "would not", tweet)
    tweet=re.sub(r"shouldn't", "should not", tweet)
    tweet=re.sub(r"can't", "can not", tweet)
    tweet=re.sub(r"isn't", "is not", tweet)
    tweet=re.sub(r"it's", "it is not", tweet)
    tweet=re.sub(r"isn't", "is not", tweet)
    tweet=re.sub(r"wasn't", "was not", tweet)
    tweet=re.sub(r"weren't", "were not", tweet)
    tweet=re.sub(r"couldn't", "could not", tweet)
    tweet=re.sub(r"won't", "will not", tweet)
    tweet=re.sub(r"W", " ", tweet)
    tweet=re.sub(r"d", " ", tweet)
    tweet=re.sub(r"s+[a-zA-Z]*s+", " ", tweet)
    tweet=re.sub(r"s+[a-zA-Z]*$", " ", tweet)
    tweet=re.sub(r"^[a-z]*s+", " ", tweet)
    tweet=re.sub(r"https", " ", tweet)
    tweet=re.strip('\ ')
    tweet=re.sub(r"http[s]*", "", tweet)
    tweet=re.sub(r"yifmqy", " ", tweet)
    tweet=re.sub(r"s+", " ", tweet)
```

## 4.7 Sentiment Analysis

We are fetching 100 tweets on iPhone from Twitter. By using our classifier model we are going to predict sentiment according to each tweet. By using NLP pipeline we cleaned redundant information and stop word for the sentiment analysis. To predict sentiment now we are using "vectorizer" and "classifier" which is "clf" object which is created from the "tfidfmodel.pickle" and "classifier.pickle". Natural Language processing is a combination with regular expression that helps to build an algorithm which can pre-process dataset with stored tweets.

```
sentiment=clf.predict(vectorizer.transform([tweet]).toarray())
print("-"+tweet+":",sentiment)

if sentiment[0]==1:
    total_pos +=1
else:
    total_neg +=1
```

Now, we use the vectorizer and the classifier to predict the sentiment corresponding to each of the tweets.

Index	Type	Size	Value
0	str	1	Embassy will make my business more successful and my tourists - happie ...
1	str	1	.@Apple approves TRAI's DND app to avoid iPhone ban in India https://t ...
2	str	1	@LFCChamp18_19 @Frxd_szn Android is literally for people that can't af ...
3	str	1	Opened Traffic Signal - Timing Inquiry request via iPhone at 28 SUN VA ...
4	str	1	Not really, with an iPhone you know what to expect but hijabis its a B ...
5	str	1	I liked a @YouTube video https://t.co/sXwIsQQSkk Honor 8X - A Budget P ...
6	str	1	NEW SPRINT TING APPLE IPHONE 7 PLUS 256GB APPLE IPHONE 7 PLUS + 256GB ...
7	str	1	9to5Mac : This week's top stories: Apple boosts iPhone trade-in value, ...
8	str	1	It's true that a brand name of iPhone's great all over the world, whic ...
9	str	1	I want AirPods/ Parfum oud C. Dior/ iPhone Xs/ 500.000frs. https://t.c ...
10	str	1	iPhone plsssss 🙏❤️ https://t.co/Kh1ky14QUi

I will have to bring both the tweet and the sentiment.



```

-embassy will make my business more successful and my tourists happier iwsdt nn gameinsight
iphonegames iphone : [1]
- apple approves trai dnd app to avoid iphone ban in india utbv ktwvs : [1]
- lfcchamp _ frxd_szn android is literally for people that can afford an iphone : [1]
-opened traffic signal timing inquiry request via iphone at sun valley bv se iupsip cmk guarantee
envhh mvhf : [1]
-not really with an iphone you know what to expect but hijabis its big surprise under all that fabric
wunqo lf : [0]
- liked youtube video sxwisqqskk honor a budget phone with iphone xs max bezels : [0]
-new sprint ting apple iphone plus gb apple iphone plus gb jet black uhtbxd uopcnwugxk : [1]
- to mac this week top stories apple boosts iphone trade in value apple music on echo more mf hydr :
[1]
-it true that brand name of iphone great all over the world which isn of urse limited to japan but do
qoondkrsfi : [1]
- want airpods parfum oud dior iphone xs frs fmkqtnbrdl : [0]
-iphone plsssss kh ky qui : [1]

```

Activate Windows  
Go to Settings to activate Windows.

Now we are going to take a sample of a tweet in order to evaluate the sentiment of the tweet. And check our classifier model whether does it give the correct sentiment among a tweet.

```
@IamKellyJoe Game of thrones is like iPhone, overrated...
```

Here we can see that '@' which is redundant information and also see many unwanted punctuation in this text to evaluating sentiment analysis. After preprocessing the tweet we can get the sentiment of this tweet.

```
iamkellyjoe game of thrones is like iphone overrated : [1]
```

Here '@' is removed from the tweet and ',', '.....' these punctuation are also removed. Our classifier model predicts the sentence as a positive sentence & gives the polarity '1'. where we can also identify the sentence actually positive. Let's see another tweet

```
Much better than an iPhone 😊 https://t.co/c1yNjARvif
```

In this tweet, an emoticon is used since we are not analyzing our tweet according to emoticon so the NLP pipeline cleans this emoticon and using many redundant words in this tweet such as "https://t.co/c1yNjARvif" which is converted into "ynjarvif" by preprocessing the tweet and this word doesn't mean anything so TF-IDF model would give zero.

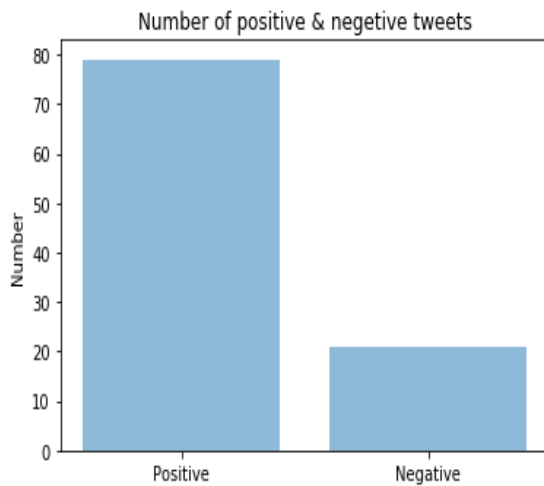
```
-much better than an iphone ynjarvif : [0]
```

Here our classifier model predicts the sentence as a negative sentence by giving 'zero' polarity. If you reduce the sentence you may see that the user wants to say about another device or object which is better than iPhone. So, the tweet actually expresses a negative statement about iPhone.

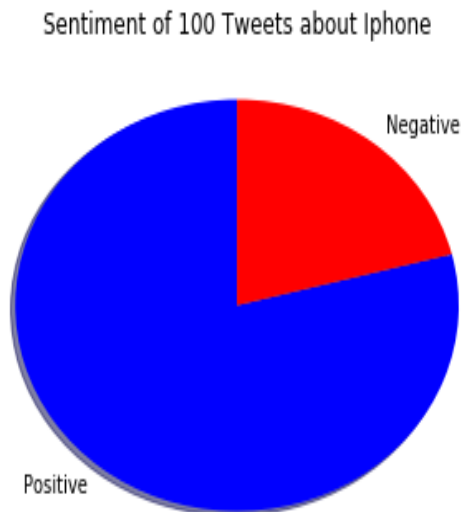
## 4.8 Plotting The Result and Accuracy of our Model

We are going to plot the number of positive tweets compared to the number of negative tweets right. Then the ratio of positive tweet & negative tweet according to fetching 100 tweet is given below by plotting a Bar Diagram

Total Positive tweet from extracting tweet= 79  
Total Negative tweet from extracting tweet= 21



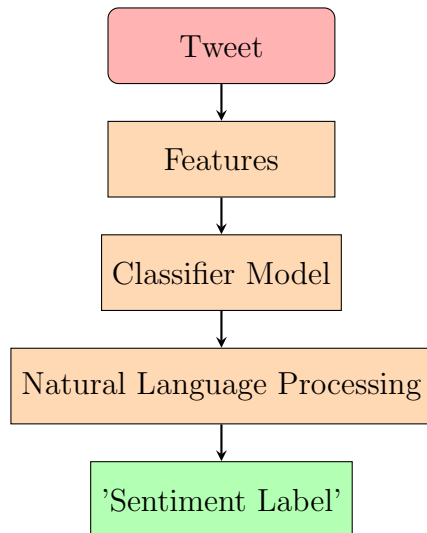
For better view we are going to plot the result in Pie-Chart:



Where "blue" color represents the positive tweet and "red" color stands for

negative tweet.

Design a architecture which is going to be classified with positive and negative tweet:



**Accuracy :**

Accuracy For min_df=1			
Max_features	min_df	max_df	Accuracy(%)
2000	1	0.1	78
2000	1	0.2	79
2000	1	0.3	81.5
2000	1	0.4	82.75
2000	1	0.5	84.75
2000	1	0.6	84.75
2000	1	0.7	84.75
2000	1	0.8	85.25
2000	1	0.9	84.75
2000	1	1.0	84.25

Here, the max\_features mean how much most frequent words can be generated as features in our histogram. min\_df means minimum document frequency where a selected word can be appeared in all of the document less than or equal given times. max\_df which is maximum document frequency in percentage where it excludes all the different words that appeared in that percentage of the document or more than.

Accuracy For min_df=2			
Max_features	min_df	max_df	Accuracy(%)
2000	2	0.1	78
2000	2	0.2	79
2000	2	0.3	80.75
2000	2	0.4	82.5
2000	2	0.5	84.75
2000	2	0.6	84.5
2000	2	0.7	84.75
2000	2	0.8	85.25
2000	2	0.9	85
2000	2	1.0	84.25
Accuracy For min_df=3			
Max_features	min_df	max_df	Accuracy(%)
2000	3	0.1	78.25
2000	3	0.2	79.25
2000	3	0.3	81.25
2000	3	0.4	82.75
2000	3	0.5	84.75
2000	3	0.6	84.75
2000	3	0.7	85
2000	3	0.8	85
2000	3	0.9	85
2000	3	1.0	84.25

According to our built model, we can get 85.25% highest accuracy. In our analysis we are taking

# Chapter 5

## Comparison

### 5.1 Comparison

In this chapter, we are discussing about our analysis which was on Twitter data by using Natural language Processing. Now we are going to represent some comparison between two device those "iPhone" series and "Samsung" note series according to our analysis.

Now, We are fetching 1000 tweet for iPhone and Samsung then Calculate the total positive and negative sentiment for both devices

```
In [1]: runfile('F:/TwitterCode/All Python Code/Twitter sentiment analysis using NLP/
comparison.py', wdir='F:/TwitterCode/All Python Code/Twitter sentiment analysis using NLP')
Total Positive tweet from extracting tweet for iPhone X seires= 726
Total Negative tweet from extracting tweet for iPhone X seires= 274
Total Positive tweet from extracting tweet for Samsung Note Series = 717
Total Negative tweet from extracting tweet for Samsung Note Series= 283
Iphone Device is more popular than Samsung Device
```

Here, we can see the iPhone phones are more popular than Samsung Phone.

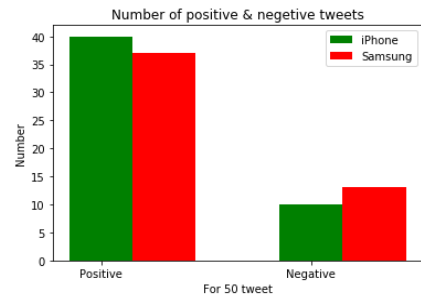
But when we are taking 100 tweet for iPhone and Samsung then let's see what's happened in the comparison:

```
In [2]: runfile('F:/TwitterCode/All Python Code/Twitter sentiment analysis using NLP/
comparison.py', wdir='F:/TwitterCode/All Python Code/Twitter sentiment analysis using NLP')
Total Positive tweet from extracting tweet for iPhone X seires= 66
Total Negative tweet from extracting tweet for iPhone X seires= 34
Total Positive tweet from extracting tweet for Samsung Note Series = 70
Total Negative tweet from extracting tweet for Samsung Note Series= 30
Samsung Device is more popular than Iphone Device
```

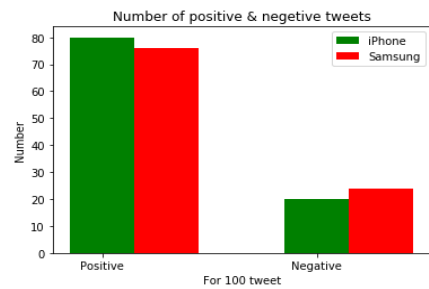
Here, we can see that the Samsung phones are more popular than iPhone. But The result doesn't create any infliction, It just gives infer about the sentiment analysis. If we work with a large number of datasets then possibly we will get a higher accurate result rather than working on less datasets. To get better understand

and visualization,we are going to plot different sentiment according to number of collecting tweet such as 50,100,500,1000 respectively.

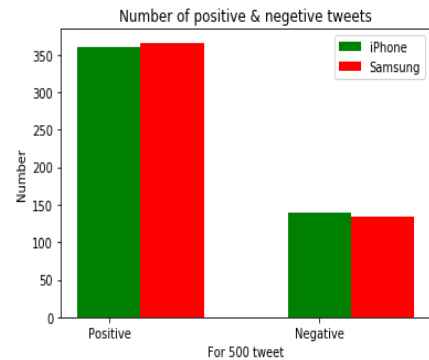
Total Positive tweet from extracting tweet for iPhone X seires= 40  
 Total Negative tweet from extracting tweet for iPhone X seires= 10  
 Total Positive tweet from extracting tweet for Samsung Note Series = 37  
 Total Negative tweet from extracting tweet for Samsung Note Series= 13  
 Iphone Device is more popular than Samsung Device



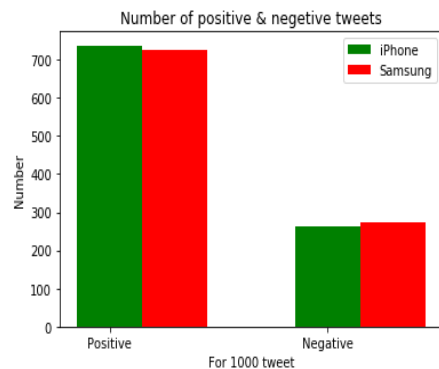
Total Positive tweet from extracting tweet for iPhone X seires= 80  
 Total Negative tweet from extracting tweet for iPhone X seires= 20  
 Total Positive tweet from extracting tweet for Samsung Note Series = 76  
 Total Negative tweet from extracting tweet for Samsung Note Series= 24  
 Iphone Device is more popular than Samsung Device



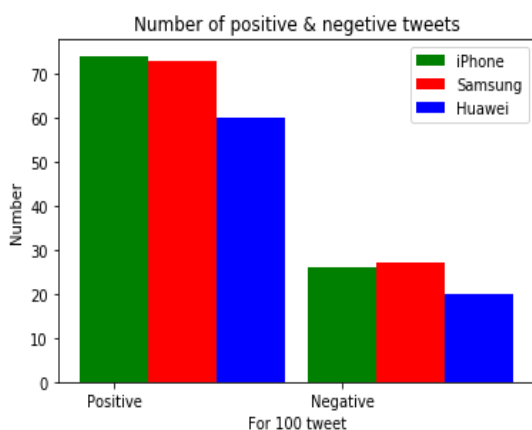
Total Positive tweet from extracting tweet for iPhone X seires= 360  
 Total Negative tweet from extracting tweet for iPhone X seires= 140  
 Total Positive tweet from extracting tweet for Samsung Note Series = 366  
 Total Negative tweet from extracting tweet for Samsung Note Series= 134  
 Samsung Device is more popular than Iphone Device



Total Positive tweet from extracting tweet for iPhone X seires= 737  
 Total Negative tweet from extracting tweet for iPhone X seires= 263  
 Total Positive tweet from extracting tweet for Samsung Note Series = 726  
 Total Negative tweet from extracting tweet for Samsung Note Series= 274  
 Iphone Device is more popular than Samsung Device



Two companies have launched many phones within a year. Some phones have taken more popularity than other company's phone and it varies phone to phone. So, it's difficult to say one company is more popular than other company because the ratio of positive and negative tweet among two phones isn't extreme. But we can say both two phone company are more popular than other phone company.



There has been a lot of process to sentiment analysis on Twitter or other social networking platform and there are also using many classifiers to predict the sentiment. At the beginning of sentiment analysis, we worked with a built-in Python library which is "Textblob". But when we able to analysis the same data by using Natural Language Processing then we have found some problem and also found some wrong prediction from "Textblob"

#### Sentiment analysis by using Python "Textblob" library

```

from textblob import TextBlob
import re

def clean_tweet(tweet):
    """
    Utility function to clean the text in a tweet by removing
    links and special characters using regex.
    """
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\w+\S+)", "", tweet).split())

def analyze_sentiment(tweet):
    """
    Utility function to classify the polarity of a tweet
    using textblob.
    """
    analysis = TextBlob(clean_tweet(tweet))
    if analysis.sentiment.polarity > 0:
        return 1
    else:
        return 0
# We create a column with the result of the analysis:
data['SA'] = np.array([ analyze_sentiment(tweet) for tweet in data['Tweets'] ])

```

## Extracting Tweet with Sentiment:

We display the updated dataframe with the new column:

	Tweets	len	SA
0	This Company Says It Can Unlock ANY Passcode-P...	85	0
1	VERY NAUGHTY! 12-year-old girl accused of pois...	116	1
2	Apple's Now Offering up to \$100 MORE for Old i...	85	1
3	4 Hidden iPhone Contacts Tricks You'll Wish Yo...	83	0
4	Man Who 'Lost' His iPhone Every Year Was Just ...	90	0
5	I replaced the battery on my iPhone 6S and it'...	134	1
6	@iPhone_News Well that depends, do you want to...	81	0
7	@iPhone_News iPhone X...	23	0
8	iPhone XR vs iPhone XS: Which should you buy t...	89	0
9	#Huawei just showed #iPhone X and #OnePlus 6T ...	109	0
10	@thenewsspeaks https://t.co/wAsAWJUpec Digitim...	140	0
11	Apple says ... https://t.co/t1fC57cZv3	36	0
12	iDrop News is giving away an iPhone XR in May....	88	1
13	iOS 12.1 Jailbreak iPhone XS Max News: NEW Cyd...	115	1
14	Barely 100 Turn Up For Ram Temple Rally In Del...	140	1
15	How to share documents and news tips with Wash...	116	0
16	News on https://t.co/GYOLESHtAm iPhone XR vs i...	122	0
17	in other news my mom finally upgraded from her...	79	0
18	Apple goes RED to mark #WorldAidsDay https://t...	73	0
19	iPhone 📱 https://t.co/148Mu060eX	32	0
20	So I filled up my daughter's car and washed it...	140	1
21	@Firefighter1613 @ijustine @iPhone_News @UrAvg...	140	1
22	I liked a @YouTube video https://t.co/SiWUCLci...	86	1
23	I liked a @YouTube video https://t.co/G3ZdzJ78...	119	1
24	Apple gives away iPhone XR to the whole audien...	140	1
25	@Firefighter1613 @ijustine @iPhone_News @UrAvg...	140	1

Here, we can see that a lot of redundant words & objects are being generated from Twitter API those have no actual meaning to give a sentiment. Now tweets are generated with a sentiment by using Natural Language processing is shown below:

-this company says it can unlock any passcode protected iphone co exgmaotb : [1]
-very naughty year old girl accused of poisoning mother for taking away iphone read more co iu pz : [1]
-apple now offering up to more for old iphone trade ins co jd0tgcynv : [1]
-hidden iphone contacts tricks you ll wish you knew sooner co dalmtxpb : [1]
-man who lost his iphone every year was just arrested in malaysia co qydzd dgls : [0]
-replaced the battery on my iphone and it pretty much running like the iphone so highly recommend co hiipj : [1]
-well that depends do you want to save money or do you want quality : [0]
-iphone : [1]
-iphone xr vs iphone xs which should you buy this holiday season co cwqzftjep : [1]
-huawei just showed iphone and oneplus how to kill the notch once and for all co lqfq : [1]
-co wasawjupec digitimes repo today that apple has enforced nd wave of order reducti co xc rd pq : [1]
-apple says co fc czv : [1]
-idrop news is giving away an iphone xr in may enter to win now co hs fuddw : [1]
-ios jailbreak iphone xs max news new cydia updates by saurik co k jnld co wjavrnaqhk : [1]
-barely turn up for ram temple rally in delhi rss expected lakhs co zlsr shared via ndtv new co lag pxv : [1]
-how to share documents and news tips with washington post journalists the washington post co urucu oral : [1]
-news on co gyoleshtam iphone xr vs iphone xs which should you buy this holiday season co qnf kgw : [1]
-in other news my mom finally upgraded from her iphone co zbg rygh : [1]
-apple goes red to mark worldaidsday co ahcpszhern via : [1]
-iphone co muo ex : [1]
-so filled up my daughter car and washed it what mess but the good news is that i solved the mystery of wh co sog wta fe : [0]

From previous extracted tweet which was polarized by "Textblob" library, we are going to take one tweet as a sample to compare:



@Firefighter1613 @ijustine @iPhone\_News  
 @UrAvgConsumer @DetroitBORG What a beautiful  
 collection of vintage Apple pr... <https://t.co/1DhHf7Imnv>

In this tweet, it's complicated to find out the actual sentence to predict whether the sentence is positive or negative because "@Firefighter1613", "@ijustine", "@UrAvgConsumer", "https://t.co/1DhHf7Imnv" those aren't containing any information or meaning. But using Natural Language Processing we have found the same sentence as "what beautiful collection of vintage apple" which is highlighted in below figure.

```
mystery of wh co sog wta fe : [0]
- am big apple fan but can use magic mous co dbqcg : [1]
- liked video co siwuclcix news manipulated by apple iphone : [0]
- liked video co zdzj gradient iphones higher iphone prices crazy hack apple news : [0]
-apple gives away iphone xr to the whole audience on the ellen show to mac general physics
laboratory gpl co zrlwevo : [1]
- what beautiful collection of vintage apple pr co ldhhf immv : [1]
-two simple tricks to make your iphone battery last all day it two clock in the afternoon
and your iphone bat co hmfxtzmvz : [1]
-win an iphone at magicred casino co keneu oiq co pycz gued : [1]
-idrop news is giving away free iphone xs max in february enter to win now co yrod1gv ta :
[0]
-apple loop poor iphone sales panic apple sudden iphone xs price cut ipad pro big mistake
via co yx plnvhbo : [0]
-snapdragon benchmark suggests androids could overpower iphone xs sma phone android news
technology co pctitkfw : [0]
-after iphone production cuts is apple dividend at risk co lqh b via : [1]
-apple loop poor iphone sales panic apple sudden iphone xs price cut ipad pro big mistake
co qvipn fo : [0]
-tech news st december gadget tech technology gadgets electronics device instagood
instatech geek techie co jukk mv : [1]
-apple pricing policies for iphone apps bought on its exclusive app store ran into trouble
monday at the supreme co qj xmrn : [1]
- microsoft beats for biggest market value co zmcjhoszs msft aapl technology software
iphone share price : [1]
-in dispute prof herbe hovenkamp sides with iphone owners seeking to sue arguing that co i
avgc : [1]
-hey here a pro tip everyone hated the notch and the no home button change let alone the no
mm jack co sjybpwajct : [1]
```

This sentence is cleaned by NLP pipeline and understandable to classify whether the sentence is positive tweet or negative. Another faced problem is given below:

```

-applemusic apple news iphone ipad ios beta how to wake up to weather forecast on your iphone
lock scr co rymwianz : [1]
-this is africa samsung nigeria tweets update using apple iphone co dz jurzz via bbc
agribusiness investor nigeria : [1]
-samsung nigeria tweets update using apple iphone co xtghqz : [1]
-samsung nigeria tweets update using apple iphone co gbtrnjmzrj : [1]
-samsung nigeria tweets update using apple iphone bbc news co rffuhpn co fiz qdlpbf : [1]
-apple will not release iphone until at least aapl via co wozmohn : [0]
-sources apple will wait until at least before offering an iphone that can connect to services
coming next co tzhvwmqfy : [0]
-apple news how to wake up to weather forecast on your iphone lock screen co ytmczdm via co re
mbtoep : [1]
-tech news samsung nigeria tweets update using apple iphone co njmftnvba via co atj : [1]
-samsung nigeria tweets update using apple iphone co iqb ytcw technology co alekdph ri : [1]
-comment real time call screening is one pixel feature d love to see on the iphone co zrmak ce
: [1]
-apsny news english samsung nigeria tweets update using apple iphone co yskjxrpok : [1]
-bbc technology samsung nigeria tweets update using apple iphone co keps ysn : [1]
-steve jobs handwritten apple specifications sheet could fetch at auction co strknfbb co qmtfr
: [0]
-apple music poised to get native android tablet interface co pivcmalx co bhi glt rw : [1]
-samsung nigeria tweets update using apple iphone co etzv awsk : [1]
-this is the case when it is better to burn out in hell samsung iphone fail epicfail
mondaymotivation co dmzovz zle : [0]
-samsung nigeria tweets update using apple iphone co sckrk izau : [1]
-techie co ptrcf samsung nigeria tweets update using apple iphone co ptrcf the account co
pctxdye gk : [1]

```

These tweets are extracting from twitter API for "iPhone" by using Natural Language Processing where the tweets are preprocessed and evaluates their sentiment according to classifier model.

```

We display the updated dataframe with the new column:

```

	Tweets	len	SA
0	AppleMusic #apple #news #iphone #ipad #ios9.3 ...	140	0
1	This is Africa - Samsung Nigeria tweets update...	130	0
2	Samsung Nigeria tweets update using Apple iPho...	72	0
3	Samsung Nigeria tweets update using Apple iPho...	72	0
4	Samsung Nigeria tweets update using Apple iPho...	107	0
5	#Apple will not release 5G #iPhone until at le...	128	0
6	Sources: Apple will wait until at least 2020 b...	139	0
7	APPLE NEWS How to Wake Up to a Weather Fo...	134	0
8	TECH NEWS Samsung Nigeria tweets update us...	123	0
9	Samsung Nigeria tweets update using Apple iPho...	108	0
10	Comment: Real-time call-screening is one Pixel...	108	1
11	Apsny News English: Samsung Nigeria tweets upd...	92	0
12	BBC Technology: Samsung Nigeria tweets update ...	88	0
13	Steve Jobs' handwritten Apple I specifications...	131	0
14	Apple Music poised to get native Android table...	105	0
15	Samsung Nigeria tweets update using Apple iPho...	72	0
16	This is the case when it is better to burn out...	131	0
17	Samsung Nigeria tweets update using Apple iPho...	72	0
18	#Techie https://t.co/59ptrcf219 Samsung Nigeri...	140	0
19	Apple Will Wait Until at Least 2020 to Release...	140	0
20	Samsung Nigeria tweets update using Apple iPho...	72	0
21	Samsung Nigeria tweets update using Apple iPho...	72	0
22	Samsung Nigeria tweets update using Apple iPho...	110	0
23	Samsung Nigeria tweets update using Apple iPho...	120	0
24	Why Apple Waiting Until 2020 for a 5G #IPHONE ...	92	0

Now the same tweets are extracting from Twitter API for "iPhone" and determine the sentiment in order to use Python "Textblob" library which library is being used to sentiment analysis. Here some tweets are highlighted from both lists of collecting tweet. Where NLP evaluate those tweet as positive whether the "Textblob" predicts as negative sentiment against those tweet. From highlighted tweets, a single tweet is taken which is "Samsung Nigeria tweets update using Apple

iPhone". This tweet is actually a positive tweet for "iPhone". According to NLP, it gives positive sentiment but the "Textblob" python library which stands for sentiment analysis that predicts the sentence as a negative tweet. Another tweet is "How to Wake Up to a Weather Forecast on Your iPhone's Lock Screen". This tweet is actually positive and NLP also evaluates the tweet as a positive tweet by given value '1' but "Textblob" predicts the wrong sentiment.

# Chapter 6

## Conclusion

### 6.1 Conclusion

The purpose of our analysis on Twitter data to show how consumer can make their decision to purchase a product with their satisfaction. Multi-national company, Entrepreneurs who want to open a new business, marketers and firms can use our proposed method accordingly Natural Language Processing to sentiment analysis on data and understand about their customers, products, services and user's need. In this paper, we represent our proposed work by a block diagram [Chapter:3]. Natural Language processing is one of the best approach to analyze sentiment. The built classifier for this work can be utilized as a data analysis tools in NLTK. The result is represented by plotting a bar diagram and also show in a Pi-chart for better visualization and understand. From the outcome of sentiment according to two mobile phone company, we have shown which mobile company is more popular than other.

### 6.2 Future Work

In future, we will work for building a Business Intelligence which can give a decision about the success or failure of a product. In our proposed work, we used a dataset which contains 2000 features and to build up classifier model we used some data as testing and others are being used as training. Since the maximum features of our dataset is 2000 so the training & testing datasets are static. If any new word comes to our model to predict sentiment then it may confused. So we will build an algorithm which can build their classifier model dynamically. If any new words come, the model will train the classifier by itself. The calculation of sentiment analysis can be more weighted by using the internal attributes of a tweet & its posterior such as their social status, job, designation, salary, previous history.

# Chapter 7

## Reference

1. <https://wersm.com/how-much-data-is-generated-every-minute-on-social-media/>
2. <https://help.twitter.com/en>  
<https://twitter.com/BarackObama>
4. [https://www.researchgate.net/publication/50378498\\_Anew\\_ANEW\\_evaluation\\_of\\_a\\_word\\_list\\_for\\_sentiment](https://www.researchgate.net/publication/50378498_Anew_ANEW_evaluation_of_a_word_list_for_sentiment)
5. <http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html>
6. [https://www.idosi.org/wasj/wasj35\(1\)17/7.pdf](https://www.idosi.org/wasj/wasj35(1)17/7.pdf)
7. [https://projekter.aau.dk/projekter/files/239450162/Twitter\\_Data\\_Mining\\_Master\\_Thesis\\_Holub\\_Final.pdf](https://projekter.aau.dk/projekter/files/239450162/Twitter_Data_Mining_Master_Thesis_Holub_Final.pdf)
8. <http://www.cs.cornell.edu/people/pabo/movie-review-data/>
9. <https://apps.twitter.com/app/>
10. <https://skymind.ai/wiki/natural-language-processing-nlp>
11. <https://github.com/SergiuTripon/stat-nlp-twitter-sentiment-analysis/blob/0d251163c548bd842c7>
12. [https://github.com/pedrobalage/python\\_natural\\_language\\_processing\\_portuguese/find/master](https://github.com/pedrobalage/python_natural_language_processing_portuguese/find/master)
13. <https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/>
14. <https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis>
16. <http://www.storybench.org/sentiment-analysis-of-you-guessed-it-donald-trumps-tweets/>
17. <https://dev.to/rodolfoferro/sentiment-analysis-on-trumpss-tweets-using-python->
18. <https://textblob.readthedocs.io/en/dev/classifiers.html>