

# **Sentiment Analysis and Opinion Mining Using NLP**

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENT FOR THE  
AWARD OF THE DEGREE OF

**Master of Technology  
Information System**

Under the guidance of

**Ms. Anamika Chauhan**

**(Department of Information Technology)**

Delhi Technological University

Submitted By

**Rajinder Singh Negi**

(Roll No. 2K19/ISY/20)



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Shahabad Daultpur, Main Bawana Road, Delhi-110042

June 2021

## DECLARATION


I hereby declare that the thesis work entitled “**Sentiment Analysis and opinion Mining Using NLP**” which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of degree of Master of Technology (Information System) is a bonafide report of Major Project-II carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Place: Delhi

Date:31/10/2021

Name- Rajinder Singh Negi

Roll No. 2K19/ISY/20

A photograph of a handwritten signature in black ink on a light-colored surface. The signature is written in a cursive style and reads "Rajinder Singh Negi". A horizontal line is drawn underneath the signature.

## CERTIFICATE

This is to certify that Project Report entitled “**Sentiment Analysis and opinion Mining Using NLP**” submitted by Rajinder Singh Negi (Roll no. 2K19/ISY/20) in partial fulfillment of the requirement for the award of degree Master of Technology (Information System) is a record of the original work carried out by him under my supervision.



Place: Delhi

Date:

**SUPERVISOR**

Ms. Anamika Chauhan

(Assistant professor)

Department of Information Technology

Delhi Technological University Bawana Road, Delhi -110042

## **ACKNOWLEDGEMENT**

I am very thankful to **Ms. Anamika Chauhan** (Assistant Professor, Department of Information Technology) and all the faculty members of the Department of Information Technology of DTU. They all provided us with immense support and guidance for the project.

I would also like to express my gratitude to the university for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

**Rajinder Singh Negi**

**Roll No. 2K19/ISY/20**

**M. Tech. (Information System)**

**Delhi Technological University**

## ABSTRACT

Social Media has become an important platform for expressing and consuming information. A customer opinion is a product and service who buy and used or had experience with that and it is Most Important for customer feedback on various product and service. Here our aim is build a Recommendation System using Natural Languages processing on Twitter Data. Each Tweet we will do required analysis on particular tweet independently. By Looking at Single Sentence (Tweets) we could not find the Sentiment of whole sentence for this need to use deep learning Neural Network for Required Analysis. Word Sentiment is useful but it cannot express the meaning of longer phrases. Therefore Sentiment detection required a richer supervised learning and more powerful model of composition

## TABLE OF CONTENTS

a) Candidate's Declaration.....	i
b) Certificate.....	ii
c) Acknowledgement .....	iii
d) Abstract.....	iv
e) Table of Contents .....	v
f) List of Tables .....	viii
g) List of Figures .....	ix
1. Introduction.....	1
1.1. <a href="#">What is Sentiment Analysis</a> .....	<a href="#">2</a>
1.2. <a href="#">Sentiment Analysis Classification</a> .....	<a href="#">3</a>
1.2.1. <a href="#">Machine learning techniques</a> .....	<a href="#">4</a>
1.2.1.1. <a href="#">Supervised learning</a> .....	<a href="#">4</a>
1.2.1.1.1. Naïve Bayes .....	5
1.2.1.1.2. <a href="#">Decision Tree</a> .....	<a href="#">9</a>
1.2.1.2. <a href="#">Unsupervised learning</a> .....	<a href="#">9</a>
1.2.2. <a href="#">Reinforcement learning</a> .....	<a href="#">10</a>
2. Literature Survey .....	18
3. Problem statement.....	21
3.1. <a href="#">Objective</a> .....	<a href="#">23</a>
3.2. <a href="#">PoS Tagging</a> .....	<a href="#">25</a>
3.3. <a href="#">Twitter dataset</a> .....	<a href="#">26</a>
3.4. <a href="#">Sentence weightage</a> .....	<a href="#">27</a>
3.5. <a href="#">Hashtag</a> .....	<a href="#">27</a>
3.6. <a href="#">Abbreviations and Redundant/Repeated letters</a> .....	<a href="#">28</a>

3.7.	<a href="#">Twitter API</a>	29
3.8.	<a href="#">Concept of pagination</a>	30
3.9.	<a href="#">Types of Streaming API</a>	30
3.10.	<a href="#">Data extraction through Twitter AP</a>	30
3.11.	<a href="#">Pre-processing of Twitter data</a>	31
3.12	<a href="#">Tweets collected from twitter</a>	36
3.13	<a href="#">Weekly wise tweet collected</a>	37
3.13.1	<a href="#">Precision</a>	38
3.13.2	<a href="#">Recall</a>	38
3.13.3	<a href="#">Accuracy</a>	38
3.14	<a href="#">Results of classifier for twitter data</a>	39
3.15	<a href="#">Results of four weeks of twitter data</a>	41
4	Conclusion and future scope	47
5	References	48

## LIST OF TABLES

Table 3.1 Weekly wise report of twitter data .....	37
Table 3.2 Results report of first week.....	42
Table 3.4 Results report of second week.....	43
Table 3.5 Results report of third week .....	44
Table 3.6 Results report of fourth week .....	45



## LIST OF FIGURES

Figure 1.1 Sentiment analysis Techniques.....	3
Figure 1.2 Parsing Natural Language Sentences.....	7
Figure 1. Methodology.....	7
Figure 1.4 No of Review count.....	8
Figure 1.6 flow diagram of proposed methodology .....	11
Figure 1.7 Rest API Working with Twitter .....	13
Figure 3.1 Streaming API working with Twitter... ..	25
Figure 3.2 Tweets collected from twitter using Twitter API .....	36
Figure 3.3 performance of classifiers.....	41
Figure 3.4 Weekly wise report of classified data .....	46

# CHAPTER 1: INTRODUCTION

The importance of Natural Language Processing to make computers understandable for natural language. but this is not an easy task. Computers Generally understand structured data like database tables but human texts, human languages, and voice are unstructured data, and for the computer main challenge to understand unstructured data. that tells the need for Natural Language Processing.

There is a lot of NLP data out there in different formats, and it's a lot easier to comprehend and process it if the computer can understand and analyse it. By training the models in various ways in line with predicted results. Humans have been writing for a long time; there is a massive amount of material available, and it would be fantastic if a computer could comprehend it. There are several difficulties in comprehending the precise meaning of a statement or correctly predicting the thoughts.

Recommendation System mainly Focuses on determining the people's opinion from the website. And Recent trends of the web encourage users to contribute and Express their point of view through e.g Blogs post, videos, social networking blogs, etc These the platform provides us important information that we are interested in analyzing. Opinion –mining system analysis who written the opinion, what is being commented by people.

## 1.1 WHY STUDY OF MACHINE LEARNING:-

1. Email Spam Filter.
2. Hand Written digit Recognition
3. Weather Forecasting
4. Movie Recommendation
5. Sentiment Analysis / Opinion Mining.

## 1.2 What is Machine Learning?

-Machine Learning is a gives machine the able to learn without being programmed.

With in the field of data analytics. Machine learning is a method used to devise complex algorithms that will predict, this is known as “predicative Analytics”.

Predicative models are needed in science, engineering, business, astronomy, biology, finance, web etc.

### Sentiment Analysis :-

Suppose you wish to purchase a thing. in the before buying a product You're looking for reviews, such as what other customers have to say about it. You check for comments, such as what other customers have to say about that particular product, whether positive or negative, and you personally examine it by looking at their comments. Consider how the company analyzed what their customers are thinking about their product level. In most cases, they don't have one or more customers. They do have a sizable customer base. So, what will they do? Here we use Sentiment analysis.

### Sentiment Analysis Classification:-

Sentiment analysis has different type of class. There are two types of this.

There are three main levels of classification: sentence level, document level, and aspect level. Polarities can be categorized into three categories based on sentiment analysis: positive, neutral, and negative.

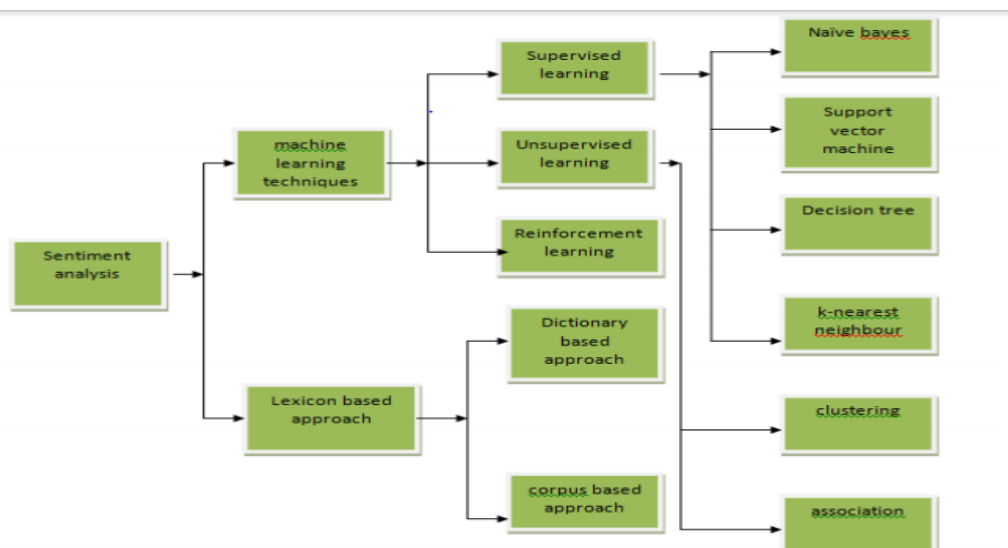


Figure 1.1. Sentiment Classification Techniques

Sentiment analysis is a broad term that encompasses a variety of methodologies. It is the most popular field of study. It is distinguished by two types: machine learning-based approaches and lexicon-based approaches. The lexicon method focuses on negative and positive terms and is divided into two types: dictionary-based and corpus-based.

## **1.2 Machine Learning techniques**

Machine learning is used to categorize the text classification problem in sentiment analysis. Training data records are utilized to train a model, which are then used to determine the predict model without level. Each record is assigned to different classes.

When we feed the model a new unlabeled record, it will label the dataset into distinct classes. Positive, negative, and neutral courses are the three sorts of classes.

### **1.2.1 Supervised learning techniques**

**Naïve Bayes (NB)** To accomplish classification tasks, a probabilistic machine learning model known as a Naive Bayes classifier is used. At the core of the classifier is the Bayes theorem.

Naive Bayes for text classification: the probability of classifying 'd' document as class 'c' given 'd' is

$$p(C/D)=p(C\cap D)/p(D)$$

**For a document d' and a class c.**

**posterior probability. Prior x Likelihood**

$$P(y | x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}$$

so according to mutually independent assumption:

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y),$$

for each value of i this function behaves:

$$P(y | x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)}$$

Joint model could be expressed as:

$$P(y | x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i | y)$$

$$\Downarrow$$

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y),$$

### posterior probability. Prior x Likelihood

The multinomial naive Bayes classifier operates as follows: To begin, we compute  $P(c|d)$  for each class  $c$  in  $C$ , which is the likelihood of returning class  $c_i$  if our observation is  $d$ . The maximum of  $P(c_1|d)$ ,  $P(c_2|d)$ ,...,  $P(c_n|d)$  is then found. consists of "our best guess for the proper class".

$$\hat{c} = \max_{c \in C} \Pr(c | d)$$

$$\Pr(c | d) = \frac{\Pr(c) \Pr(d | c)}{\Pr(d)}$$

$$\hat{c} = \max_{c \in C} \Pr(c) \Pr(d | c)$$

**Assumptions:**-1. Bag of words - position of words in document doesn't matter.

2 conditional independence -  $P(x_j|c)$  are independent given the class  $c$ .

$$\hat{c} = \max_{c \in C} \Pr(c) \Pr(f_1, f_2, \dots, f_n | c)$$

$$\begin{aligned} \Pr(c_i) &= \frac{\text{number of docs of class } c}{\text{total number of docs in training dataset}} \\ &= \frac{N_c}{N_{docs}} \end{aligned}$$

We use a collection of documents labelled  $(d_1, c_1), \dots, (d_n, c_n)$  by classes to train our model. Using the training dataset, we compute  $P(c)$ , the prior probability, and  $P(f_1|c)P(f_2|c)\dots P(f_n|c)$ , the likelihood probability, for each class. In our approach, we store these variables in a data structure. To account for zero probability, we use add-one Laplace smoothing.

$$\begin{aligned} \Pr(w_i|c) &= \frac{\text{count}(w_i, D_c) + 1}{\sum_{w' \in V} (\text{count}(w', D_c) + 1)} \\ &= \frac{\text{count}(w_i, D_c) + 1}{\sum_{w' \in V} (\text{count}(w', D_c)) + |V|} \\ &= \frac{\text{count}(w_i, D_c) + 1}{\sum_{d \in D_c} (\text{len}(d)) + |V|} \quad \text{more intuitive sum} \end{aligned}$$

## (DT) Decision Tree

A decision tree is a supervised learning approach that uses a classification algorithm. It's a graphical representation of most of the options for making a choice based on a set of criteria.

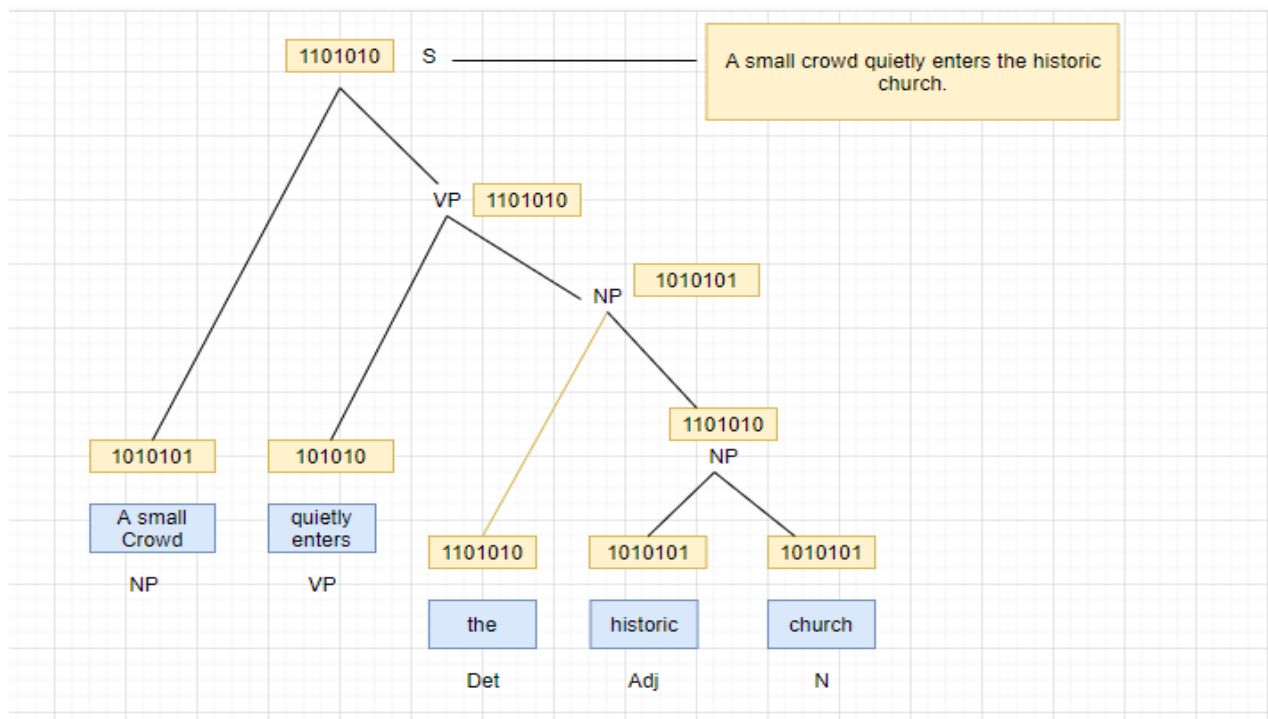


Fig 1.2 Parsing Natural Language Sentences

The main Challenges is are

- Positive sentiment words ( Example :-good , great ,awesome, wonderful etc) by this we can't say that complete sentence is positive sentiments and vice versa.

Example:-If I can find a good car in the showroom I will buy it. (Good positive sentiments but whole sentences have neutral sentiment).

- The +ve or -ve word may have opposite orientation in different Application Domains

Example:- This phone sucks.( -ve sentiment ).

This vacuum cleaner really sucks (+ve Sentiment ).

- Word Embedding: Converting each word into a vector and conversion should be such a way that two similar words should have minimum distance.
- Here NLP parse the sentence based upon grammatical rule.

### Unsupervised learning

It's a machine-learning method. It does not use data that has not been categorised or tagged. It enables the algorithm to operate on such data without the need for human intervention. The goal of unsupervised learning is to categorise unsorted data into groups based on patterns.

There is no human direction offered in unsupervised learning, which implies the computer does not get any training data. As a result, the machine is able to discover the underlying structure in unlabeled data on its own. Clustering and association are two different types of unsupervised learning.

# Chapter 2

## Literature Survey

Recommendation System mainly Focuses on determining the people's opinion from the website. And Recent trends of the web encourages users to contribute and Express their point of view through e .g Blogs post, videos , social networking blogs etc These platform provides us important information that we are interested for analysis. Opining –mining system analysis who written the opinion ,what is being commented by people.

Opinion mining determine the polarity (+ve, -ve) and Polarity scale (weakly Positive, mildly positive., highly negative etc) .

**Current research is focusing on:**

- Reducing the human effort need to analyze large content.
- Semantic analysis through corpus (bags) of words with known sentiment for sentiment classification
- Political Debates (Polls)
- Product Review

### **2.1 Here we want to improve the accuracy for Sentiment analysis, for that we are going through various method**

- 1 Statistical way to analysis the sentiment as mentioned in [1].
2. By studying various learning (Navies, Maximum Entropy, Support vector machine) and studying the accuracy of each learning as mentioned in [2].
3. By improving the pre-processing (removing the irreverent thing and selecting the future from each tweet)

Preprocessing- Tokenization, Case-Folding, Removing Stopping Word, Stemming and lemmatization.

4. By improving the Word Embedding (the word or document which are closely related they have same vector (array) with minimum distance).



5. By improving the POS (part of speech tagging) as mentioned in [3].

**2.2 Here we Don't require any dataset** Because we will create our dataset by its own through sentiment analysis. By select the future from each twists and storing them into mysql and computation time will not increase because we will doing sentiment analysis in multithreading environment. For each twits we will create a thread for that analysis.

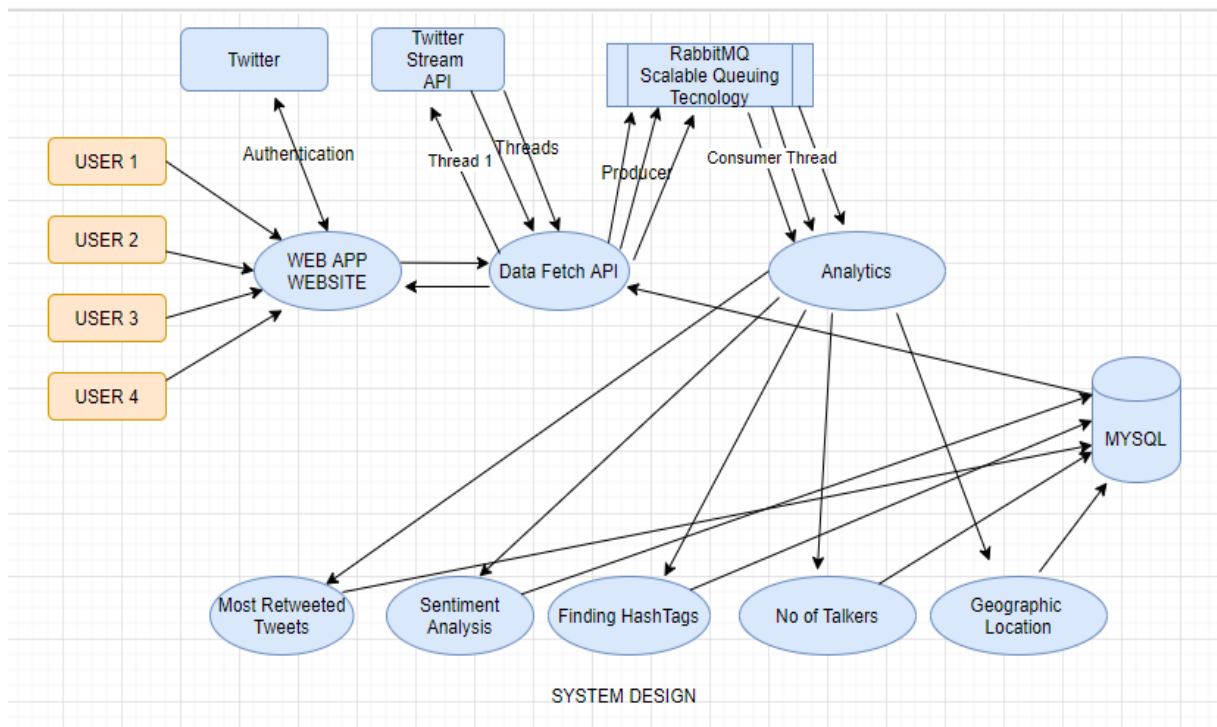


Fig.2.1 How dataset will be created for Sentiment Analysis

# STATISTICAL AND SENTIMENT ANALYSIS OF CONSUMER PRODUCT REVIEWS

Analysis a large set of online Review for mobile phone is Conduct. Here they are not only classify the text but also scale the sentiments between positive to negative. This method of categorising reviews is useful for evaluating products and assisting consumers and product owners in making better decisions.

The words 'amazing' and 'degrade' have positive and negative polarity, respectively. If they use negative terms like "not," the attitudes shift dramatically. However, statements containing negative words may sometimes exacerbate rather than change the polarity.

We calculated sentence level sentiment orientation of the reviews into ten different sentiments using Statistical and Sentiment Analysis of Consumer Product Reviews (SACP),

**Collecting from Amazon.com, the e-commerce behemoth. Over 400,000 evaluations for roughly 4500 mobile phones make up the data collection. As previously stated, it has six characteristics.**

<b>Feature</b>	<b>Description</b>
Product Name	Model name of the mobile phone.
Brand Name	Brand associated with the mobile phone.
Price	Price of the mobile set in dollars.
Rating	User rating between 1 to 5.
Reviews	User reviews provided for every mobile phone.
Review Votes	Number of people who found the review helpful.

Table [2.1]:-Feature include in the Data Set

# STATISTICAL ANALYSIS OF DATA SET

## 1. Number of Review Counts by Brand

Conclusion:- Among all the other brands, Samsung, BLU, and Apple have the most customers.

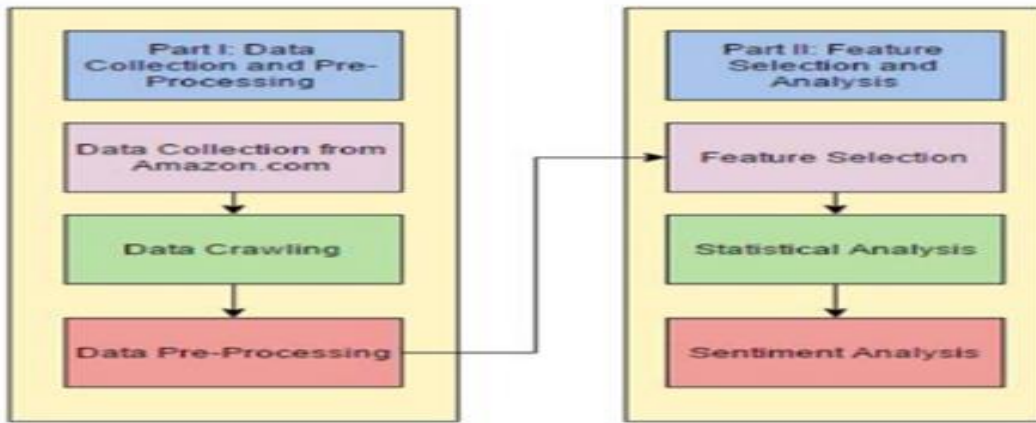


Figure 2.2: Methodology Used

# STATISTICAL ANALYSIS OF DATA SET

## 1. Number of Review Counts by Brand

Conclusion:- Among all the other brands, Samsung, BLU, and Apple have the most customers.

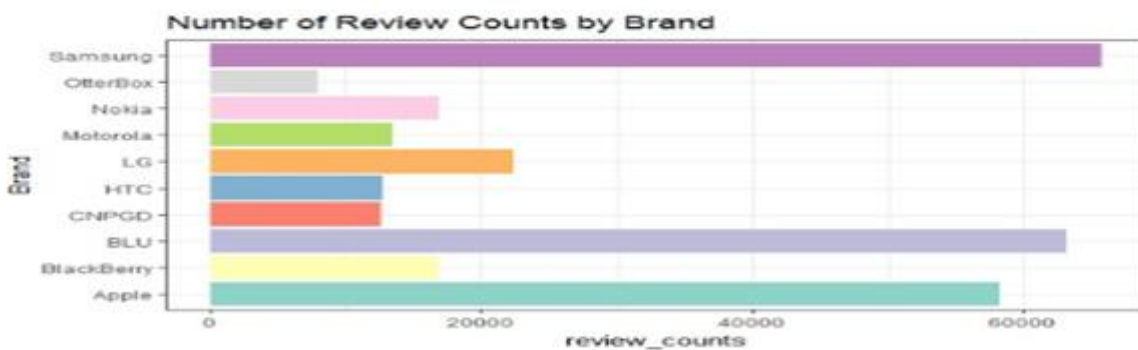


Fig 2.3: Number of Review Counts by Brand

2. Rating Distribution by Brand found that Samsung is the most popular brand, while OtterBox has the lowest rating. Here's a stack bar graph that shows how many ratings a brand has received.

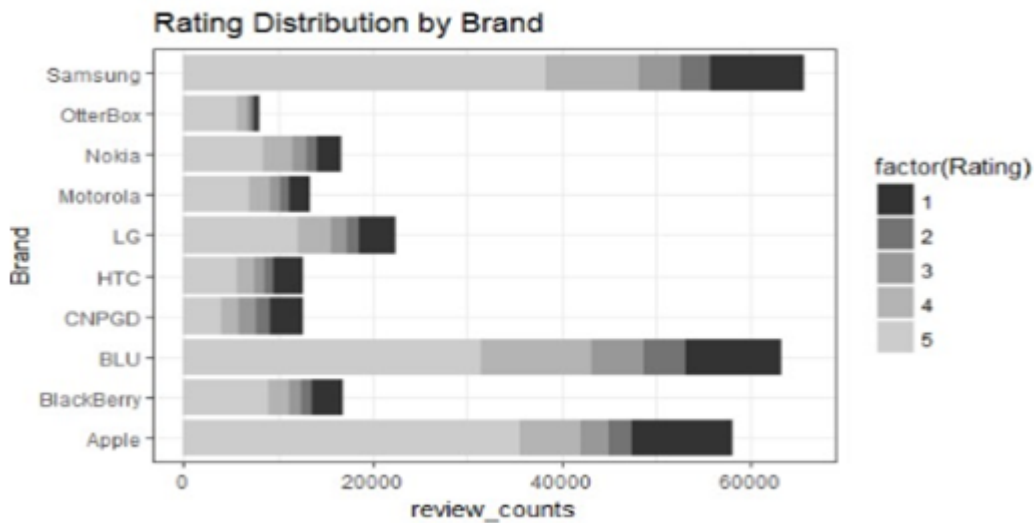
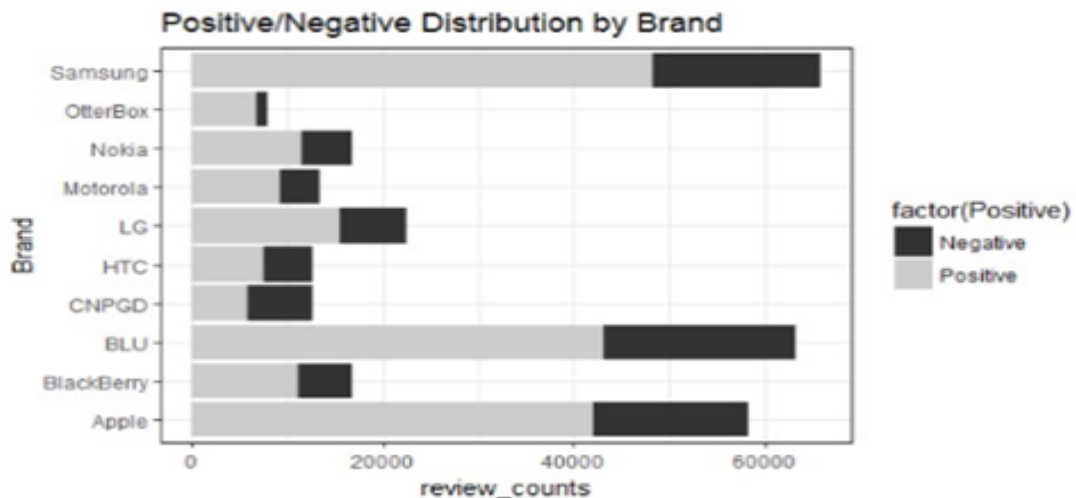


Fig 2.4: Rating Distribution by Brand

3. Reviews, both positive and negative Branded Distribution. The positive and negative feelings associated with each brand are used to classify the reviews. This study will reveal client purchasing habits. Customer choose the brand have maximum positive review. It seems that Brand Samsung has highest positive review among all the brands.



Fig

2.5:-Positive and Negative Reviews Distribution by Brand

#### 4 Review Length:-

4.1 Review Length and Product Rating: Indicates whether or not the review is influenced by the product rating.

4.2 Review Length and Product Price: This demonstrates that when the price rises, the length of the reviews does not.

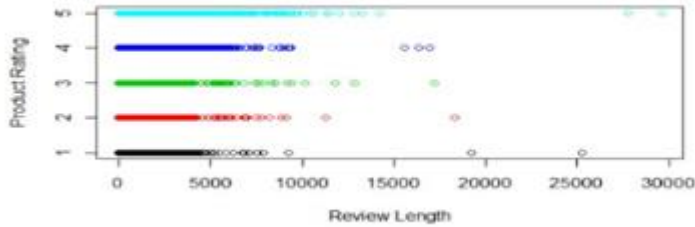


Fig 2.6:- Review Length and Product Rating

3 Price and Rating:- It Tell about Higher will be the price of the product, higher the expectations of the consumers and better should be the quality . High priced products attract higher ratings illustrating higher satisfaction among buyers of expensive products.

4 Word Cloud:-tell us what frequent word used for that product. It tell that what consumer think about the product.



Fig 2.7:- Word Cloud for Samsung Reviews



Fig 2.8:- Word Cloud for BLU Reviews



Fig 2.9:-Word Cloud for Apple Reviews

Based Upon this Statically data the sentiment analysis are Shown

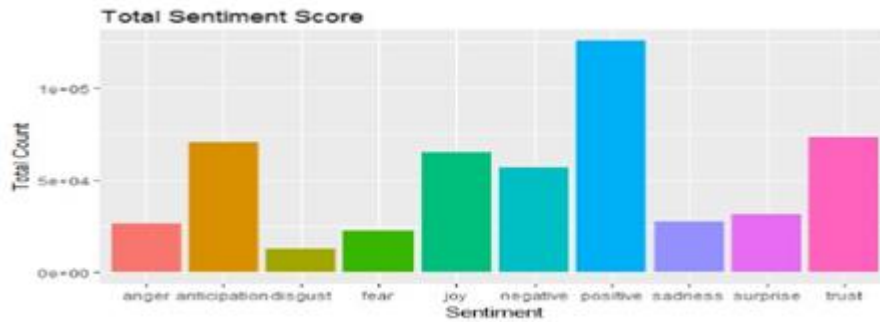


Fig 3.0:-Sentiment Analysis of Samsung Reviews

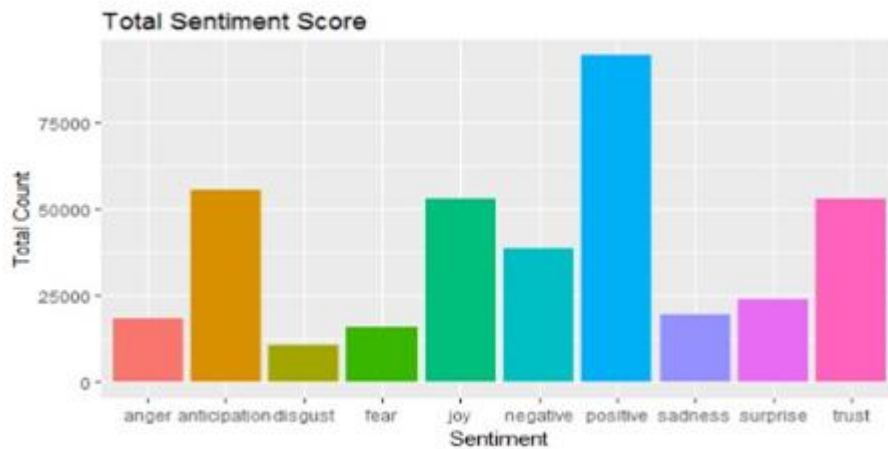


Fig 3.1:-Sentiment Analysis of Apple Reviews

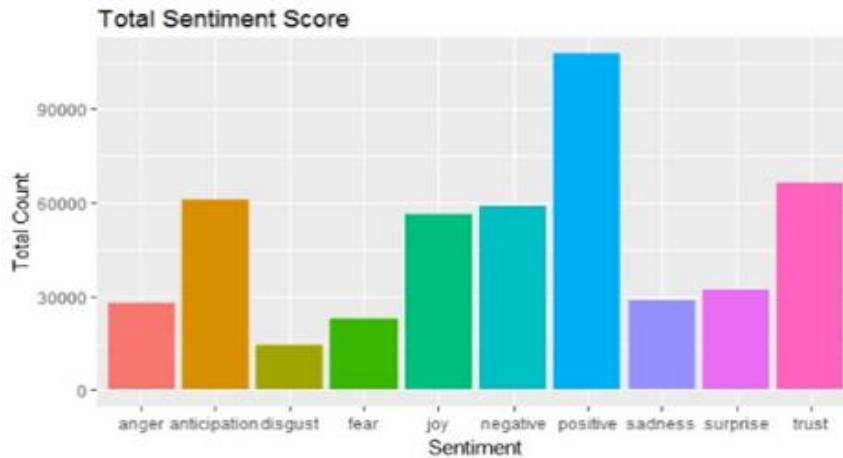


Fig 3.2:-Sentiment Analysis of BLU Review

## **Machine Learning and Semantic Analysis Approaches to Sentiment Analysis of Twitter Data**

Analyzing the data in terms of the number of tweets where the sentiment is very unstructured, either good or -ve, or somewhere in the middle. Following the preparation of the dataset (tweets), we extracted the feature vector from the dataset, and then used the ML based classification algorithm to the feature vector list: -Nave Bayed, Maximum Entropy, and Support Vector Machines

FIG. 2. FLOW DIAGRAM OF THE PROPOSED METHODOLOGY



Fig. 2. Flow Diagram of the proposed methodology

TABLE I. NAIVE BAYESIAN CLASSIFICATION MEASUREMENTS

Performance Measures (%)	
Positive Recall	91.2
Negative Recall	85.4
Positive Precision	49.3
Negative Precision	39.3

TABLE II. MAXIMUM ENTROPY MEASUREMENTS

Performance Measures (%)	
Positive Recall	86.1
Negative Recall	80.0
Positive Precision	40.4
Negative Precision	33.6

- **Precision:** This give you how many times they were actually positive. And you predict something positive,
- **Recall:** This give you how many times you predicted correctly out of actual positive data.

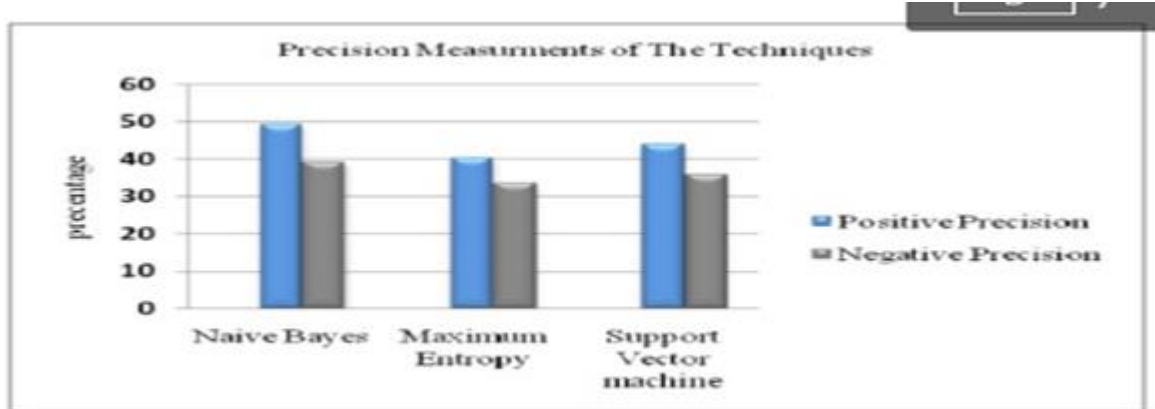


Fig 3.3:-Measurements of positive and Negative precision of the techniques



**Conclusion:**-Maximum Entropy and SVM give by the naïve bayes technique.

## Unsupervised Twitter Sentiment Classification

The method involves NLP technique, unsupervised rule based classification and WSD(Word Sense Disambiguation). This method will find the polarity of the tweets either positive or negative or in between. Overall tweet polarity for tweet is taken by rule based classifier.

### 2.3 Word Sense Disambiguation

The word “sick ‘ used in negative sense “ I feel very sick today” whereas it is used with a positive sense “Your vacuum cleaner sick very well” This algo assigned a word meaning that is most related to a given set of words

WSD Process determine the best<word ,POS –tag ,sense>. The word with highest confidence score is considered as best match

And Negative and positive Sentiment scores are determine by Senti WordNet , based on <word, POS-tag> matching .And Obtained information is represented as follows : <word, POS, sense, Positive Score, Negative Score >.

### 2.4 Tweet Polarity Classification

This phase we determine the overall tweet polarity by rule based classifier

$w$  – token from tweet having sentiment/emotion score

$score(w, p)$  – positive sentiment score of token  $w$

$score(w, n)$  – negative sentiment score of token  $w$

$score(w, o)$  – objective (neutral) sentiment score of token  $w$

$$Pos = \sum_{w \in tweet} score(w, p)$$

if (score(w,p) > 0 and score(w,p) > score(w,n))

$$Neg = \sum_{w \in tweet} score(w, n)$$

if (score(w,n) > 0 and score(w,n) > score(w,p))

$$Obj = \sum_{w \in tweet} score(w, o)$$

if (score(w,o) > 0 and score(w,p)=0 and score(w,n)=0)

$$PosCnt = \sum_{w \in tweet} w$$

if (score(w,p) > 0 and score(w,p) > score(w,n))

$$NegCnt = \sum_{w \in tweet} w$$

if (score(w,n) > 0 and score(w,n) > score(w,p))

$$ObjCnt = \sum_{w \in tweet} w$$

if (score(w,o) > 0 and score(w,p)=0 and score(w,n)=0)

$$TokCnt = \sum w$$

if (score(w,p) > 0 or score(w,n)=0 or score(w,o)>0)

PR=PosCnt/TokCnt (Positive count ratio)

NR=NegCnt/TokCnt (Negative count ratio)

OR=ObjCnt/TokCnt (Objective /neutral count ratio)

Table 1: Classification conditions for each polarity class: positive, negative and neutral.

Polarity	Classification condition
positive	$\frac{Pos}{Neg} \geq \frac{3}{2}, \frac{Pos}{Obj} \geq \frac{3}{2}, PR > NR, PR > OR$
negative	$\frac{Neg}{Pos} \geq \frac{3}{2}, \frac{Neg}{Obj} \geq \frac{3}{2}, NR > PR, NR > OR$
neutral	$\frac{Obj}{Pos} \geq \frac{3}{2}, \frac{Obj}{Neg} \geq \frac{3}{2}, OR > PR, OR > NR$

## 2.5 Evaluation Datasets

### Stanford Twitter Sentiment Test Set (STS-Test)

	A	B	C	D	E	F
1	0	1467810369	Mon Apr 06 22:19:45	NO_QUERY	_TheSpecialOne_	@switchfoot http://
2	0	1467810672	Mon Apr 06 22:19:49	NO_QUERY	scotthamilton	is upset that he can't
3	0	1467810917	Mon Apr 06 22:19:53	NO_QUERY	mattycus	@Kenichan I dived n
4	0	1467811184	Mon Apr 06 22:19:57	NO_QUERY	ElleCTF	my whole body feel:
5	0	1467811193	Mon Apr 06 22:19:57	NO_QUERY	Karoli	@nationwideclass n
6	0	1467811372	Mon Apr 06 22:20:00	NO_QUERY	joy_wolf	@Kwesidei not the v
7	0	1467811592	Mon Apr 06 22:20:03	NO_QUERY	mybirch	Need a hug
8	0	1467811594	Mon Apr 06 22:20:03	NO_QUERY	coZZ	@LOLTrish hey long
9	0	1467811795	Mon Apr 06 22:20:05	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope th
10	0	1467812025	Mon Apr 06 22:20:09	NO_QUERY	mimismo	@twittera que me m
11	0	1467812416	Mon Apr 06 22:20:16	NO_QUERY	erinx3leannexo	spring break in plain
12	0	1467812579	Mon Apr 06 22:20:17	NO_QUERY	pardonlauren	I just re-pierced my
13	0	1467812723	Mon Apr 06 22:20:19	NO_QUERY	TLeC	@caregiving I couldr
14	0	1467812771	Mon Apr 06 22:20:19	NO_QUERY	robobbierobert	@octolinz16 It it cou
15	0	1467812784	Mon Apr 06 22:20:20	NO_QUERY	bayofwolves	@smarrison i would'
16	0	1467812799	Mon Apr 06 22:20:20	NO_QUERY	HairByJess	@iamjazzyfizzle I wi
17	0	1467812964	Mon Apr 06 22:20:22	NO_QUERY	lovesongwriter	Hollis' death scene v
18	0	1467813137	Mon Apr 06 22:20:25	NO_QUERY	armotley	about to file taxes
19	0	1467813579	Mon Apr 06 22:20:31	NO_QUERY	starkissed	@LettyA ahh ive alw
20	0	1467813782	Mon Apr 06 22:20:34	NO_QUERY	gi_gi_bee	@FakerPattyPattz Ol
21	0	1467813985	Mon Apr 06 22:20:37	NO_QUERY	quanvu	@alydesigns i was oi
22	0	1467813992	Mon Apr 06 22:20:38	NO_QUERY	swinspeedx	one of my friend call
23	0	1467814119	Mon Apr 06 22:20:40	NO_QUERY	cooliodoc	@angry_barista I bak

Table 3: Accuracy, precision, recall and F-measure for the positive polarity class.

	STS-Test	STS-Gold	Sanders	SemEval
<b>Accuracy</b>	81.908	<b>83.764</b>	75.261	78.737
<b>Precision</b>	<b>81.818</b>	76.144	54.074	80.475
<b>Recall</b>	<b>91.443</b>	86.867	86.140	84.907
<b>F-positive</b>	<b>86.363</b>	81.152	66.441	82.632

**Conclusion :-**This process will find the sentiment among -ve, neutral and +ve for every tweets. Based on rule based classification .Experimentally result come this method good for complex task

# CHAPTER 3

## Twitter API

**Implementation** :- Here we will use to Create a Account for build App for Twitter data. To get the Access Token and secret key from Twitter. We will get the following things from twitter :-

- For the App
  - Consumer Key
  - Consumer Secret

For the user

- Key
- Secret

### 3.1 Twitter API:-Twitter have two API

- Rest API
- Streaming API

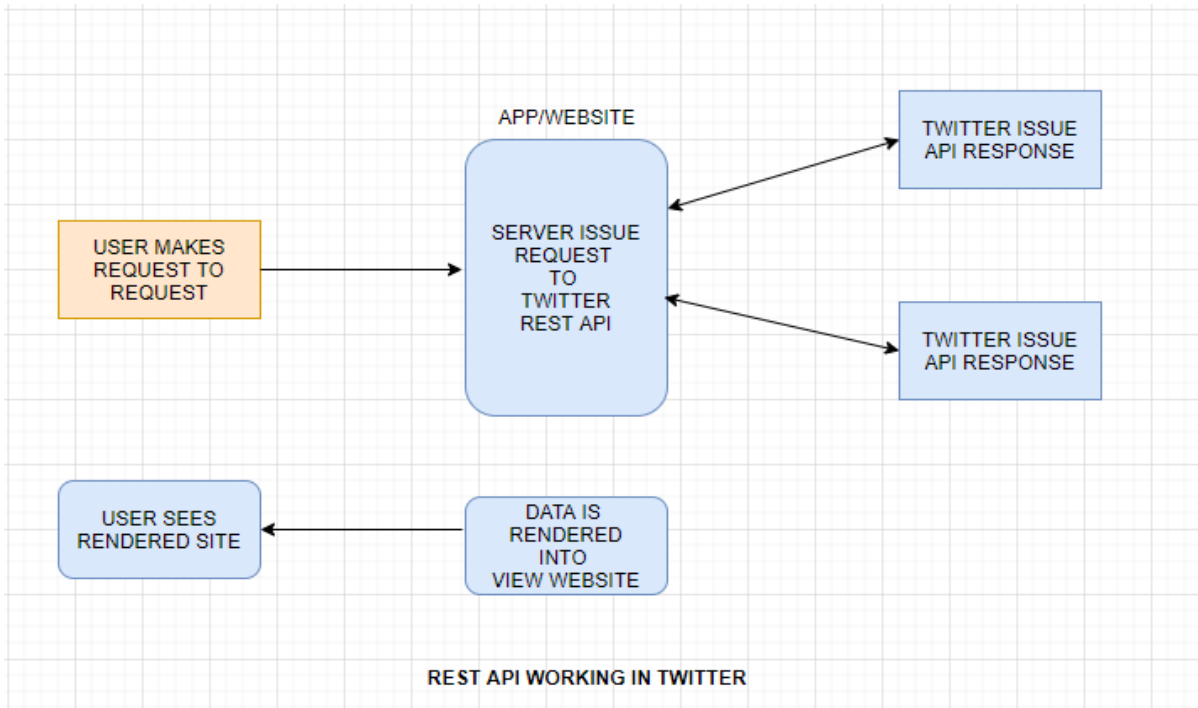


Fig 3.1

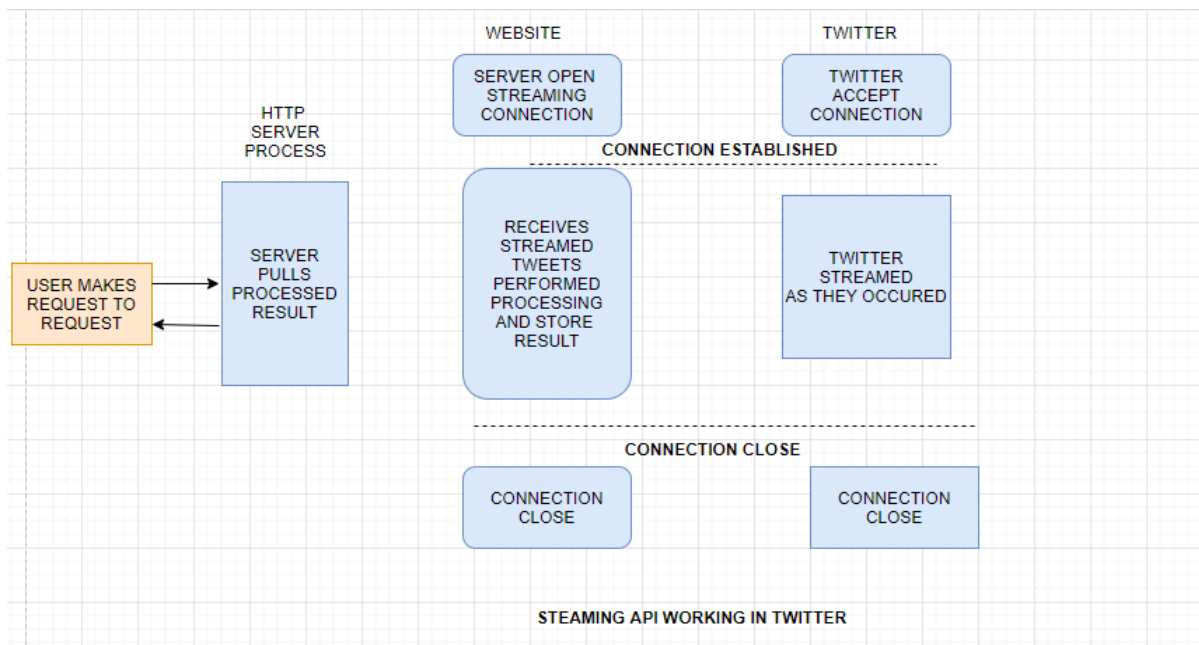


Fig.3.2

Here Both the API work in https .Difference in Twitter REST API is that it Work as Stateless manner means when we send a single request and reply then there will be no connection will made just as 1 Request and 1 Reply. Whereas in Streaming API for Request and Reply Connection is made and twists will flow as a stream and Further we will store them into our local database.

Streaming API are 3 types Public Stream , User Stream and Site Stream.

Public Stream:- whatever twists twitter getting if you want all those Twists. Here we can filter based upon topic.

User Stream:- Based upon user whoever Login into your Application ,you want to analysis Only those twists.

Site Stream:-you can put Few user in 1 group and you might want to see what all this user generally get in twists.

### Difference Between REST API And Streaming API

Both of them are going to connect the server using HTTP only but then coming to REST API no state will be maintained at server side, Everything is forgotten by server. In Streaming API once you open a stream continuously data flow in a stream. And state will be maintained.

How to get all Tweets?

## By concept of **Pagination** (Present in Twitter API Doc)

Pagination is way for requesting all of the pages.

Problem since we talk about tweet once you get fixed no of tweet in 1<sup>st</sup> page then there is chance that after you get the most recent tweet then again some might of tweets posted then how you are going to deal with that.

Let us consider a case after receiving 1<sup>st</sup> page before get 2<sup>nd</sup> page let say some tweets are newly tweeted means there is some updation. Then their will be some inconsistence among tweet .

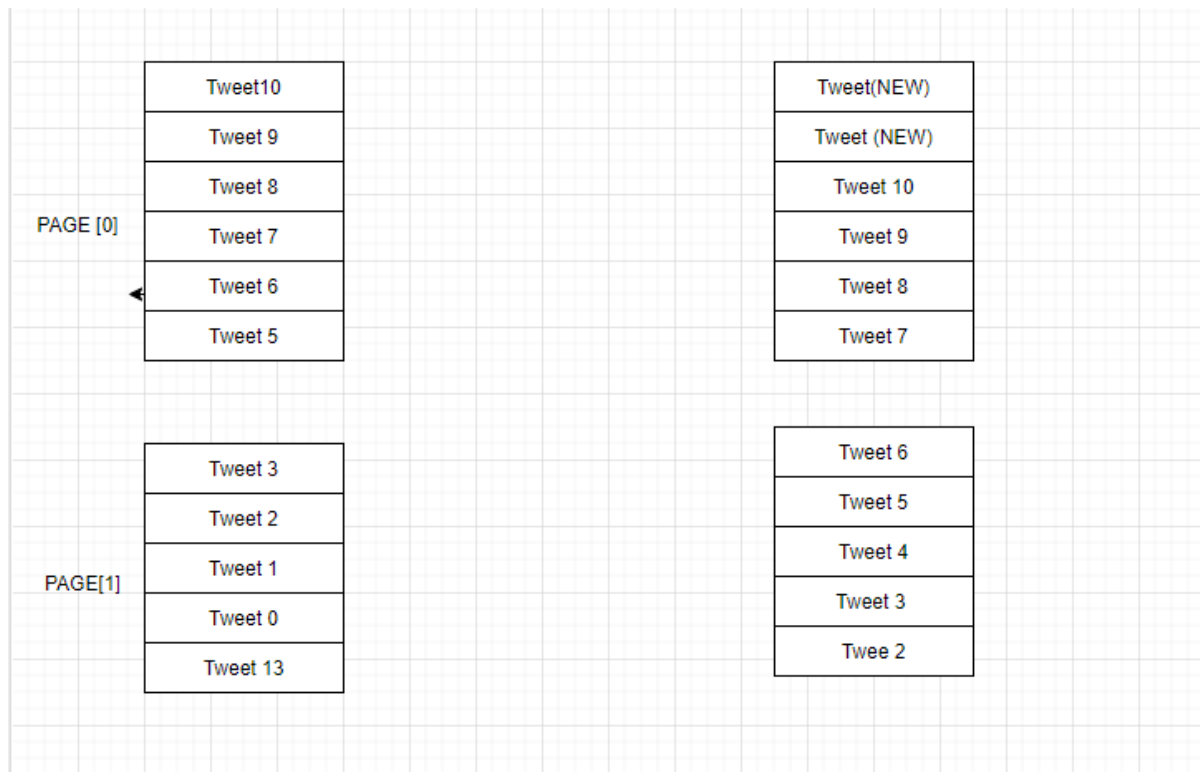


Fig 3.3

Resolve this problem by using max\_id parameter:

Whenever send request for asking for tweets you can tell server that so far we receive the tweet till this point you can send after that.

Using Since\_id to access the newly added tweets.

## TYPES OF STREAMING API METHOD

1 GET:- In get method you can send only endpoint url only Upto certain limits around 250 character..

2.POST:- In POST method has url and body in which you send the query regarding your searches and Filter the person for which you want to get twists.

SEARCH KEYWORD:- 400 SEARCH KEYWORD

And FOLLOW :-5000 FILTER

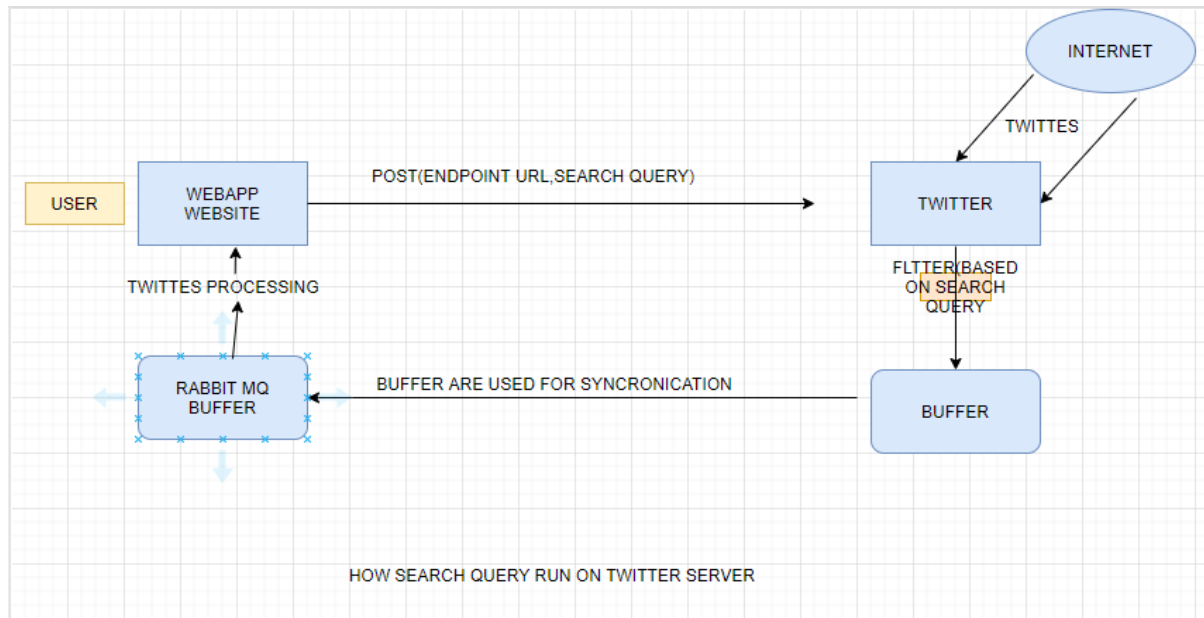


Fig.3.4

```
Untitled7.ipynb
File Edit View Insert Runtime Tools Help
Open in playground
Viewing
[ ] import os
import tweepy as tw
import pandas as pd

[ ] consumer_key= 'yEAbfrqizJ9BX0cA3wlmGaTQy'
consumer_secret= 'YmoJKxobDgtFJqrUcOswlLdz08zghHmDjG6iZLF1F7VyyQPRSu'
access_token= '852031078244077571-L112p0xYlI5RksCJ64hWk57lseuudBj'
access_token_secret= 'ledYFTU53k2XbqCHD2f14iuE2LiHdz9W3eRKZk83qVtVv'

[ ] auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)

[ ] # Post a tweet from Python
api.update_status("Look, I'm not again tweeting from #Python in my #earthanalytics class! @EarthLabCU")
# Your tweet has been posted!

Status(api=<tweepy.api.API object at 0x7f029e761be0>, _json={'created_at': 'Fri Jun 05 06:39:31 +0000 2020', 'id': 1268794562689986560, 'id_str': '1268794562689986560'})

[ ] # Define the search term and the date_since date as variables
search_words = "#wildfires"
date_since = "2018-11-16"
```

Untitled7.ipynb

File Edit View Insert Runtime Tools Help

Open in playground

```
[ ] # Define the search term and the date_since date as variables
search_words = "#wildfires"
date_since = "2018-11-16"

[ ] # Collect tweets
tweets = tw.Cursor(api.search,
                    q=search_words,
                    lang="en",
                    since=date_since).items(5)

tweets

<weeepy.cursor.ItemIterator at 0x7f029e783400>

[ ] # Collect tweets
tweets = tw.Cursor(api.search,
                    q=search_words,
                    lang="en",
                    since=date_since).items(5)

# Iterate and print tweets
for tweet in tweets:
    print(tweet.text)

RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..
RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..
RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..
```

12:42 PM  
INTL 05-06-2020

Untitled7.ipynb

File Edit View Insert Runtime Tools Help

Open in playground

```
RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..
RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..
RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..
RT @wildfiretoday: About 66 miles of Interstate 80 closed west of Salt Lake City for the #PacificFire
#UTfire #wildfires
RT @sahar_saki: Zagros forest is burning and there is no international help, so I made a pool in PayPal, donation link
https://t.co/gyWgo-

[ ] new_search = search_words + " -filter:retweets"
new_search

'#wildfires -filter:retweets'

[ ] # Collect tweets
tweets = tw.Cursor(api.search,
                    q=search_words,
                    lang="en",
                    since=date_since).items(5)

# Collect a list of tweets
[tweet.text for tweet in tweets]

["RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..",
"RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..",
"RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 🙏 the #fireal..",
"RT @wildfiretoday: About 66 miles of Interstate 80 closed west of Salt Lake City for the #PacificFire\n#UTfire #wildfires"]
```

12:43 PM  
INTL 05-06-2020



Untitled7.ipynb ☆

File Edit View Insert Runtime Tools Help

Open in playground Viewing

```

["RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 📍 the #fired...",
"RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 📍 the #fired...",
"RT @northyorkmoors: ⚠️ Despite this week's rain, there's still a high risk of #wildfires here, which could devastate wildlife 📍 the #fired...",
'RT @wildfiretoday: About 66 miles of Interstate 80 closed west of Salt Lake City for the #PacificFire\n#UTfire #wildfires',
'RT @sahar_saki: Zagros forest is burning and there is no international help, so I made a pool in PayPal, donation link\n\nhttps://t.co/gYNgo...']

[ ] tweets = tw.Cursor(api.search,
                        q=new_search,
                        lang="en",
                        since=date_since).items(5)

users_locs = [[tweet.user.screen_name, tweet.user.location] for tweet in tweets]
users_locs

[ ] [[ 'VisionsofSamad', 'Cleveland, OH'],
      ['abstractwallart', 'Kihei, HI'],
      ['iifong', 'New York, NY'],
      ['kaskadia', 'Cascadia'],
      ['FereJohn', 'SoCal']]

[ ] new_search = search_words + " -filter:retweets"
new_search

'#wildfires -filter:retweets'

[ ] tweets = tw.Cursor(api.search,
                        q=new_search,

```

Windows taskbar: 12:44 PM 05-06-2020

Untitled7.ipynb ☆

File Edit View Insert Runtime Tools Help

Open in playground Viewing

```

new_search

'#wildfires -filter:retweets'

[ ] tweets = tw.Cursor(api.search,
                        q=new_search,
                        lang="en",
                        since=date_since).items(5)

[tweet.text for tweet in tweets]

['@illhueminati #wildfires #Australia #palestine #trumpimpeachment #covid19 #BlackLivesMatter World, hold on. https://t.co/jd9712cgs8',
'California #artist Danielle Nelisse paints #abstract landscape #paintings of #daniellenelisse #wildfires | thank yo... https://t.co/YfCdcCFv5X',
'@NatHurricaneCon @disastersafety provides valuable #catastrophe resources for businesses to prepare for #hurricanes... https://t.co/0TRTWiI2eX',
'\#IllegalLogging rises in Brazil's Rondonia state, sparking concerns ahead of the #FireSeason\': "As COVID-19 spread... https://t.co/X8VdGqg25r',
'In addition, it is suggested by some environmentalists that #wildhorses can protect wild areas. forests,... https://t.co/idfwemmxXa']

tweets = tw.Cursor(api.search,
                    q=new_search,
                    lang="en",
                    since=date_since).items(5)

users_locs = [[tweet.user.screen_name, tweet.user.location] for tweet in tweets]
users_locs

[['PauCostaF', 'worldwide'],
 ['VisionsofSamad', 'Cleveland, OH'],
 ['abstractwallart', 'Kihei, HI'],

```

Windows taskbar: 12:46 PM 05-06-2020

Untitled7.ipynb

File Edit View Insert Runtime Tools Help

Open in playground

```

new_search = '#wildfires -filter:retweets'

[ ] tweets = tw.Cursor(api.search,
                        q=new_search,
                        lang="en",
                        since=date_since).items(5)

[ ] [tweet.text for tweet in tweets]

[ ] ['@illhueminati #wildfires #Australia #palestine #trumpimpeachment #covid19 #BlackLivesMatter World, hold on. https://t.co/jdg7l2cgs8',
     'California #artist Danielle Nelisse paints #abstract landscape #paintings of #daniellenelisse #wildfires | thank yo... https://t.co/YfCdcCFv5X',
     '@NatHurricaneCon @disastersafety provides valuable #catastrophe resources for businesses to prepare for #hurricanes... https://t.co/0TRTwII2eX',
     '\#Illegallogging rises in Brazil's Rondonia state, sparking concerns ahead of the #FireSeason': "As COVID-19 spread... https://t.co/X8VdGqg25r",
     'In addition, it is suggested by some environmentalists that #wildhorses can protect wild areas. forests,- https://t.co/ldfwemrXa']

[ ] tweets = tw.Cursor(api.search,
                        q=new_search,
                        lang="en",
                        since=date_since).items(5)

[ ] users_locs = [[tweet.user.screen_name, tweet.user.location] for tweet in tweets]
[ ] users_locs

[ ] [['PauCostaF', 'worldwide'],
     ['VisionsofSamad', 'Cleveland, OH'],
     ['abstractwallart', 'Kihei, HI']]

```

Windows taskbar: ENG 12:46 PM INTL 05-06-2020

Untitled7.ipynb

File Edit View Insert Runtime Tools Help

Open in playground

```

[['PauCostaF', 'worldwide'],
 ['VisionsofSamad', 'Cleveland, OH'],
 ['abstractwallart', 'Kihei, HI'],
 ['iiiorg', 'New York, NY'],
 ['kaskadia', 'cascadia']]

[ ] tweet_text = pd.DataFrame(data=users_locs,
                             columns=['user', 'location'])
[ ] tweet_text

[ ]
   user  location
0  PauCostaF  worldwide
1  VisionsofSamad  Cleveland, OH
2  abstractwallart  Kihei, HI
3  iiiorg  New York, NY
4  kaskadia  Cascadia

[ ] new_search = "climate+change -filter:retweets"

[ ] tweets = tw.Cursor(api.search,
                        q=new_search,
                        lang="en",
                        since=date_since).items(1000)

```

Windows taskbar: ENG 12:48 PM INTL 05-06-2020

Untitled7.ipynb

File Edit View Insert Runtime Tools Help

Open in playground

4 kaskadia Cascadia

```
[ ] new_search = "climate+change -filter:retweets"

tweets = tw.Cursor(api.search,
                    q=new_search,
                    lang="en",
                    since='2018-04-23').items(1000)

all_tweets = [tweet.text for tweet in tweets]
all_tweets[:5]
```

```
[ 'Looking at Climate Change and its disastrous repercussions,I am reminded f this statement frm Ban Ki Moon(UN Sec Ge... https://t.co/aHjsoQ5nQj',
  'Today, on World Environment Day, we are excited to tell you about Climate Sunday, a way to get your church involved... https://t.co/g7QLe3xfwi',
  '@Anthony_Potts@Glinner Case in point: climate change. It's now essentially illegal to discuss it.',
  'Happy #WorldEnvironmentDay! 🌍🌱\n\nGreater Manchester is leading the way on the green agenda. That's why we're one of... https://t.co/Rih4h8nXj',
  'Our work in environment & climate change adaptation addresses the interlinked challenges of disaster risk, sustaina... https://t.co/kdASPU6Lzn']
```

Windows taskbar: ENG 12:49 PM, INTL 05-06-2020

## Classify each tweet as Positive, Negative or Neutral.

We follow 3 major steps:

- Authorize the client .
- Do a GET Request to Twitter for extracting the tweets for query what client generated.
- Parse it and classify into 3 group neutral, -ve or +ve.

```
Positive tweets percentage: 32.857142857142854 %
Negative tweets percentage: 25.714285714285715 %
Neutral tweets percentage: 41.42857142857143 % \
```

#### Positive tweets:

```
RT @ssoff: Donald Trump and David Perdue have subjected us to one of the most stunning displays of incompetence and corruption in American...
@realDonaldTrump Said who? Trump? My favorite statement though. Abraham Lincoln, Franklin D. Roosevelt, and George... https://t.co/aEjWSP8ZDy
RT @PalmerReport: It's one in the morning and Donald Trump just tweeted "Melissa is great!" At this point he's totally gone.
RT @SethAbramson: FUN FACT: Donald Trump got his ass handed to him in the 2020 election. He got blown out like a chump. His performance was...
RT @Svon31: Not sure how many times the Charlatan-in-Chief wants to lose Georgia but apparently he's a glutton for punishment!
```

#### #TrumpFail...

```
RT @Olivianuzzi: In Georgia, Donald Trump just slipped up and admitted he lost the election before quickly correcting himself and saying he...
RT @donwinslow: Donald Trump is humiliating himself (and our country) on @FoxNews right now.
@msjanebond007 Robert MAXWELL (THE MIRROR) also has a certain similarity to Rupert MURDOCH, AC, KCSG (THE SUN / NEW... https://t.co/6C34Eg2aJP
RT @meridithmcgraw: VALDOSTA, Georgia – President Donald Trump on Saturday night tried on a new role: campaigning for someone not named Don...
Donald Trump faces trial over rent-fraud scheme affecting 14,000 New York tenants https://t.co/FmT1b1Wz9
```

#### Negative tweets:

```
RT @AdamSchiff: Let's make this clear:
```

```
Donald Trump lost. By a lot.
```

```
And in defeat, Trump could care less about a worsening pandemic and s...
```

```
RT @PalmerReport: To give you an idea of just how poorly Trump's Georgia rally went for the Republicans, his fans drowned out Perdue and Lo...
RT @TheTweetOfGod: Say what you will about Donald Trump, but you have to admit: he's a piece of shit.
@TeamTrump @realDonaldTrump "You can't con people, at least not for long. You can create excitement, you can do won... https://t.co/gudvjey4Hu
RT @johnpavlovitz: You owe an apology to every child who has to spend their formative years in an America defined by:
fear of the other,
an...
```

```
RT @staffand: Donald Trump kallar allt som han tycker olika för "Fake News". DW kallar det som man anser för "Det här är FAKTA - Ingen åsik...
@BuckForColorado @realDonaldTrump "You can't con people, at least not for long. You can create excitement, you can... https://t.co/EEPI7karXL
Turns out, Donald Trump is worse at cheating than he was at being president, which was worse than being a businessm... https://t.co/vSbLLahG50
RT @MartyMcL: EXCL: Donald Trump's Turnberry hotel was paid nearly £25,000 by the Secret Service for business trips by Eric Trump & his wi...
@kayleighmcenany @realDonaldTrump @MELANIATRUMP "You can't con people, at least not for long. You can create excite... https://t.co/FQ9Rf2T2Cg
```

Steps:-

- Tokenize(tweet) ,
- Eliminate stop words from the tokens.(stop words those which commonly used irrelevant in text analysis like I, is, a etc.)
- Take token and do POS(Part of Speech) Tagging like adjectives, adverbs, etc.
- Pass tokens to a **sentiment classifier** so that we can classifies the tweet by give polarity between -1.0 to 1.0 .
- -ve and +ve features are extracted from each tweets.
- Training data consists of labelled -ve and +ve . This data is trained on a NBC(Naïve Bayes Classifier).
- Then, polarity division as:

```
If . polarity > 0:  
    return 'positive'  
  
else if. polarity < 0:  
    return 'negative'  
  
else:  
    return 'neutral'  
,
```

- Finally it return parsed tweets . Then we do statistical analysis on these tweets.we tried to find the percentage of +ve, -ve and neutral tweets about a query.

## Sentiment Analysis Using VADER

**VADER(Valence Aware Dictionary and Sentiment Reasoner ) :**

VADER not only tells Whether it is -ve and +ve through score but also tells us about how much -ve or +ve sentiment is.

```

1st statement :
Overall sentiment dictionary is : {'neg': 0.165, 'neu': 0.588, 'pos': 0.247, 'compound': 0.5267}
sentence was rated as 16.5 % Negative
sentence was rated as 58.8 % Neutral
sentence was rated as 24.7 % Positive
Sentence Overall Rated As Positive

2nd Statement :
Overall sentiment dictionary is : {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
sentence was rated as 0.0 % Negative
sentence was rated as 100.0 % Neutral
sentence was rated as 0.0 % Positive
Sentence Overall Rated As Neutral

3rd Statement :
Overall sentiment dictionary is : {'neg': 0.437, 'neu': 0.563, 'pos': 0.0, 'compound': -0.4767}
sentence was rated as 43.7 % Negative
sentence was rated as 56.3 % Neutral
sentence was rated as 0.0 % Positive
Sentence Overall Rated As Negative

```

```

sentiment_dict = sid_obj.polarity_scores(sentence)

print("Overall sentiment dictionary is :", sentiment_dict)
print("sentence was rated as ", sentiment_dict['neg']*100, "% Negative")
print("sentence was rated as ", sentiment_dict['neu']*100, "% Neutral")
print("sentence was rated as ", sentiment_dict['pos']*100, "% Positive")

print("Sentence Overall Rated As", end = " ")

# decide sentiment as positive, negative and neutral
if sentiment_dict['compound'] >= 0.05 :
    print("Positive")

elif sentiment_dict['compound'] <= - 0.05 :
    print("Negative")

else :
    print("Neutral")

```

### Positive Sentiment

(score>=0.05)

### Neutral Sentiment :

(score>-0.05)and(score<0.05)

### Negative Sentiment :

(score <= -0.05)

# NLP analysis of Reviews

## Step 1: Text Cleaning or Pre-processing

- Remove Punctuations, Numbers.
- **Stemming:** Take roots of the word.

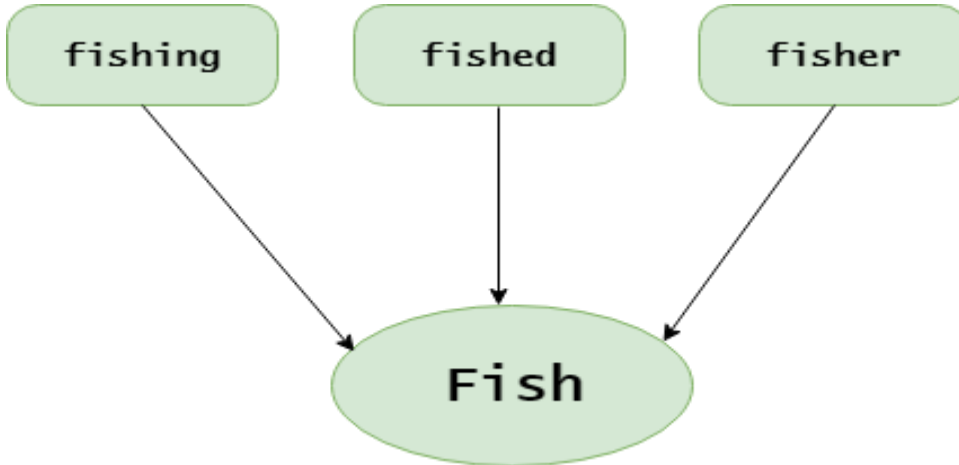


Fig 3.5 :-Before and after applying above code (reviews => before, corpus => after)

Index	Review
684	Dann good steak.
685	Total brunch fall.
686	Prices are very reasonable, flavors are spot on, the sauce is home mad.
687	The decor is nice, and the piano music soundtrack is pleasant.
688	The steak was amazing...rge fillet relleno was the best seafood plate _
689	Good food , good service .
690	It was absolutely amazing.
691	I probably won't be back, to be honest.
692	will definitely be back!
693	The sergeant pepper beef sandwich with auju sauce is an excellent sand..
694	Hawaiian Breeze, Mango Magic, and Pineapple Delight are the smoothies _
695	Went for lunch - service was slow.
696	We had so much to say about the place before we walked in that he expe..
697	I was mortified.
698	Needless to say, we will never be back here again.
699	Anyways, The food was definitely not filling at all, and for the price..
700	The chips that came out were dripping with grease, and mostly not edib..
701	I wasn't really impressed with Strip Steak.
702	Have been going since 2007 and every meal has been awesome!!
703	Our server was very nice and attentive as were the other serving staff.

Index	Type	Size	Value
684	str	1	dann good steak
685	str	1	total brunch fall
686	str	1	price reason flavor spot sauc hone made slow drench mayo
687	str	1	decor nice piano music soundtrack pleasant
688	str	1	steak amaz rge fillet relleno best seafood plate ever
689	str	1	good food good servic
690	str	1	absolut amaz
691	str	1	probabl back honest
692	str	1	definit back
693	str	1	sergeant pepper beef sandwich auju sauc excel sandwich well
694	str	1	hawaiian breez mango magic pineapple delight smoothi tri far good
695	str	1	went lunch servic slow
696	str	1	much say place walk expect amaz quickli disappoint
697	str	1	mortift
698	str	1	needless say never back
699	str	1	anyway food definit fill price pay expect
700	str	1	chip came drip greas mostli edibl
701	str	1	realli impress strip steak
702	str	1	go sinc everi meal aweson
703	str	1	our server was very nice and attentive as were the other serving staff

## Step 2: Creating the BOW(bag of word)

good	food	servic	dam	steak
1	0	0	1	1
2	1	1	0	0

### Dataset Description:

- Column divided through \t
- First column is related to reviews.
- In second column, 1 for positive and 0 for negative.

Review LikedWow... Loved this place. (1)Crust is not good. (0)Not tasty and the texture was just nasty. (0)Stopped by during the late May bank holiday off Rick Steve recommendation and loved it. (1)The selection on the menu was great and so were the prices. (1)Now I am getting angry and I want my damn pho. (0)Honeslty it didn't taste THAT fresh.) (0)The potatoes were like rubber and you could tell they had been made up ahead of time being kept under a warmer. (0)The fries were great too. (1)A great touch. (1)Service was very prompt. (1)Would not go back. (0)The cashier had no

**Step 3** Divide into Test and Traning set

75/25 via "test\_size"

**Step 4** : Predictive Model (random\_forest)

**Step 5** : Pridict Final Results.

```
array([[0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1,
        1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1,
        0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0,
        0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1,
        1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0,
        0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1,
        1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1,
        0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0,
        0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0,
        0, 1, 0, 0, 1, 1, 1, 0])
```



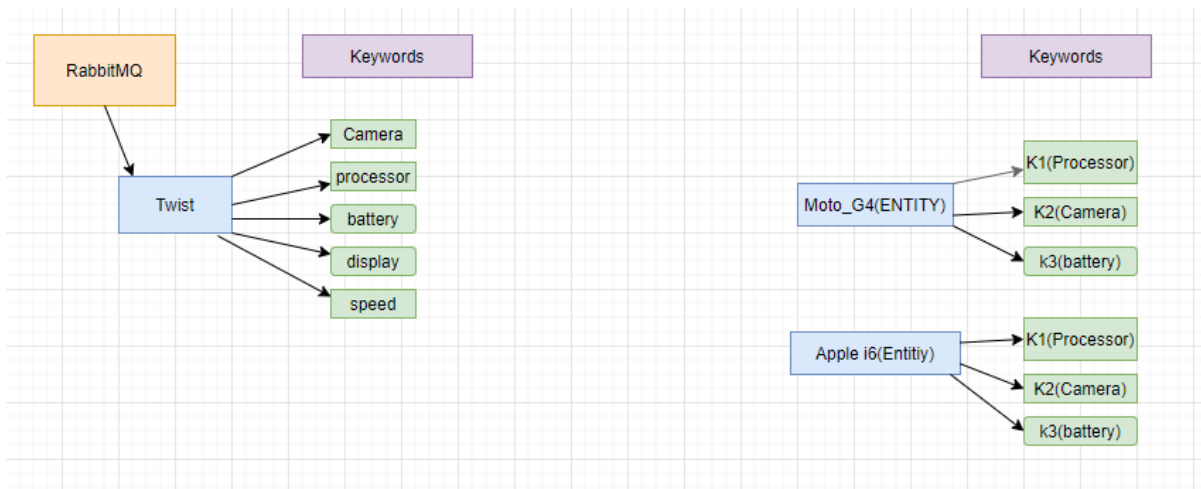
Accuracy with random forest was 72%.

**Step 6:** Accuracy determine by Confusion\_Matrix

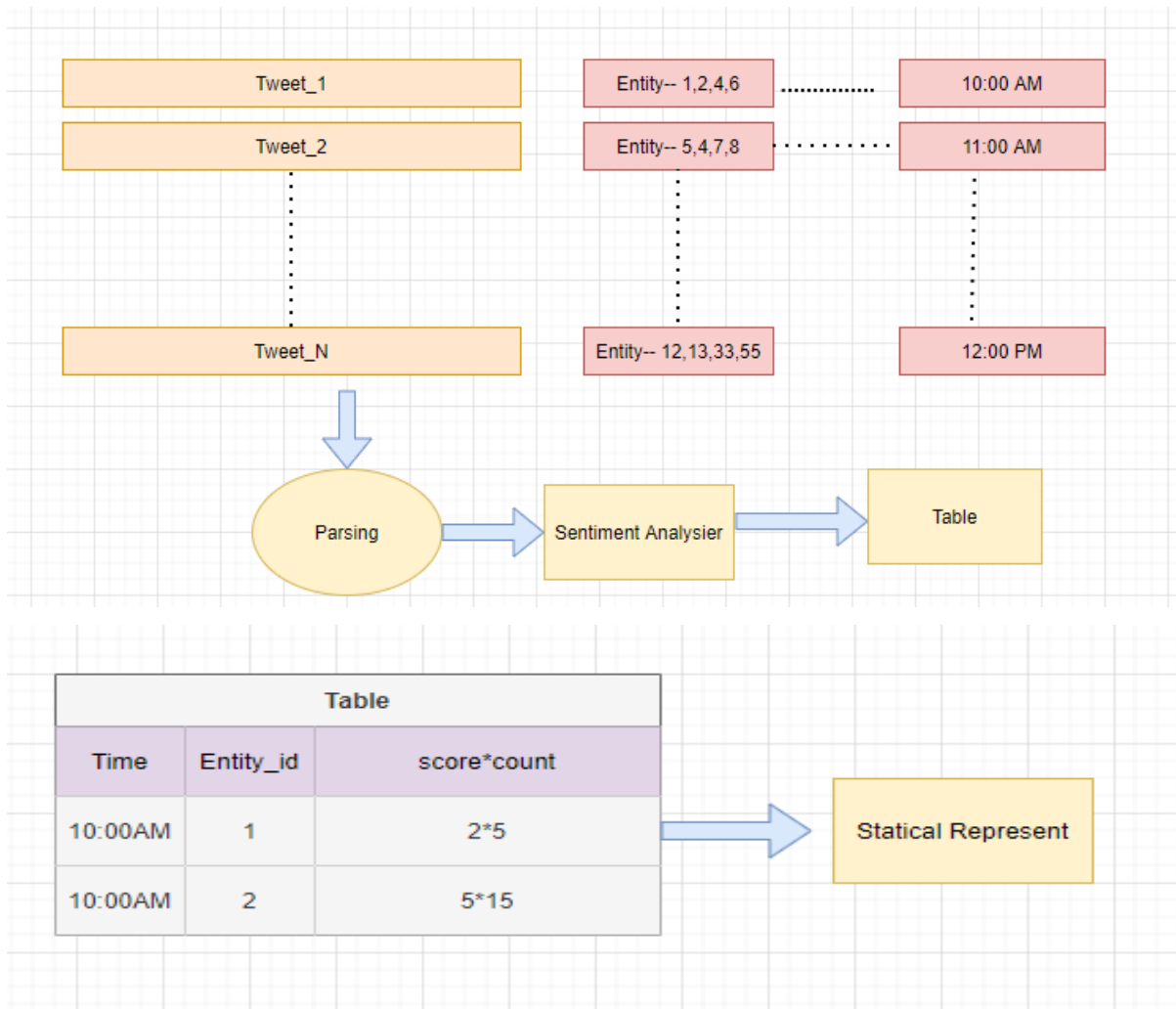
X_test = 200	Pridicted NO	Prediced YES
Actual NO	TRUE POSITIVES	FALSE NEGATIVES
Actual YES	FALSE POSITIVES	TRUE NEGATIVES

```
array([[ 77,  36],  
       [ 34, 103]])
```

### 3.6 How Sentiment of a Tweet converted into Statistical Data







Weekly wise tweet collected

Week1		Week2		Week3		Week4	
Date	No. of tweets	Date	No. of tweets	Date	No. of tweets	Date	No. of tweets
15/08/21	369	22/08/21	04	29/08/21	40	15/08/21	14
16/08/21	272	23/08/21	08	30/08/21	33	15/08/21	40
17/08/21	150	24/08/21	20	31/08/21	21	15/08/21	77
18/08/21	16	25/08/21	10	01/08/21	18	15/08/21	29
19/08/21	41	26/08/21	23	02/08/21	09	15/08/21	07
20/08/21	16	27/08/21	53	03/08/21	51	15/08/21	09
21/08/21	06	28/08/21	44	04/08/21	09	15/08/21	33
						15/08/21	18

Table 3.1

This table is comprise of four weeks twitter data. This data fetch from twitter through twitter API . This data is unstructured data Therefore we will do sentiment analysis in this twitter data. Then we will perform classification on this data. We classify this twitter data into 3 broad class such as positive ,negative neutral.

### 3.7 Performance metrics of sentiment classification:

Generally, to measure the performance of classifier we use standards such as precision, accuracy ,recall. Accuracy is dependent on two measures.

#### 3.7.1 Precision

We do division of total no right classifying positive observation by the total no predict positive observation.

$$\text{Precision} = \frac{tp}{tp + fp}$$

#### 3.7.2 Recall

Recall is a ratio between the right classifying positive observation to total no. of positive observation. high recall shows that class(division) is rightly classified.

$$\text{Recall} = \frac{tp}{tp + fn}$$

#### 3.7.3 Accuracy

To Know which model gives a better result, it is necessary to know the accuracy.

Accuracy of any model find through as:

Where,

Tp =true\_positive

Tn =true\_negative

Fn =false\_negative

Fp = false\_positive

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fn} + \text{tn} + \text{fp}}$$

### 3.8 Results of classifier for twitter data:

In this thesis, different classifier has tested on same dataset some gave best performance on the bases of precision, recall and accuracy. Data fetched from the twitter dataset. Below is the performance of each classifier.

- Naïve bayes(NB):

```
NB accuracy: 0.6341463414634146  
NB Precision: 0.6271929824561403  
NB Recall: 0.89375
```

Support vector machine(SVM)-

---

```
SVM accuracy: 0.7896341463414634  
SVM Precision: 0.8531468531468531  
SVM Recall: 0.7625
```

Decision tree(DT):

---

```
DT accuracy: 0.7957317073170732  
DT Precision: 0.8301886792452831  
DT Recall: 0.825
```

---

K-nearest neighbor(k-NN):

---

```
KNN accuracy: 0.7073170731707317  
KNN Precision: 0.8557692307692307  
KNN Recall: 0.55625
```

---



Table 3.2

- **First week results of classified data**

		Week 1			
Date	Negative Tweet	Neutral tweet	Positive tweet	Total tweet	
15/08/21	15	42	312	369	
16/08/21	10	104	168	272	
17/08/21	02	63	85	150	
18/08/21	03	09	04	16	
19/08/21	02	21	18	41	
20/08/21	02	13	01	16	
21/08/21	02	03	01	06	
Total	36	255	589	870	

Table 3.3

**Second week results of classified data**

	Week 2			
Date	Negative Tweet	Neutral tweet	Positive tweet	Total tweet
22/08/21	02	01	01	04
23/08/21	03	04	01	08
24/08/21	05	11	04	20
25/08/21	02	03	05	10
26/08/21	05	08	10	23
27/08/21	10	22	21	53
Total	27	49	42	118

**Third week results of classified data**

	Week 3			
Date	Negative Tweet	Neutral tweet	Positive tweet	Total tweet
29/08/21	09	11	20	40
30/08/21	05	12	16	33
31/08/21	05	13	03	21
01/08/21	03	03	12	18
02/08/21	04	03	02	09
03/08/21	08	23	20	51
21/08/21	03	04	02	09
Total	37	69	76	181

Table 3.4

**Fourth week results of classified data**

		Week4		
Date	Negative tweets	Neutral tweets	Positive tweets	Total tweets
05/09/21	04	04	05	13
06/09/21	12	15	13	40
07/09/21	03	23	51	77
08/09/21	06	13	10	29
09/09/21	03	01	03	07
10/09/21	03	01	05	09
11/09/21	07	03	23	33
12/09/21	04	02	12	18
13/09/21	05	03	07	15
14/09/21	10	10	20	40
Total	57	75	149	281

Table 3.5

Statistics for this week is shown clear that We got 281 total number of tweets. Out of which 149 positive tweets. And also we got 75 and 57 neutral and negative tweets respectively.

In this above week there is hike in positive number of tweet compare to second week and third week.

And people have given more negative tweet compare to second week and third week. 75 people have given their mixed opinion.

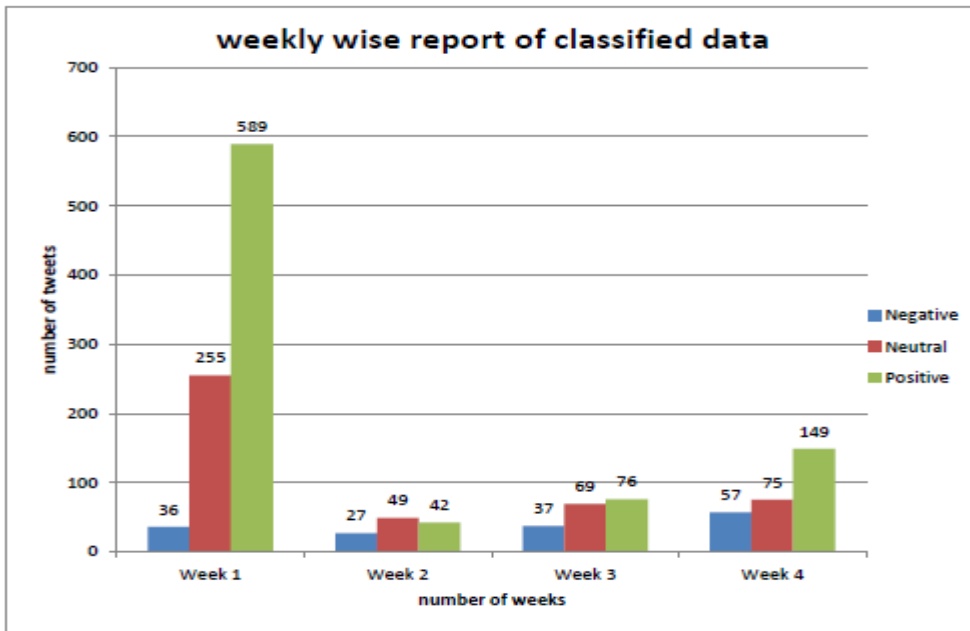


Table 3.6 weekly wise report of classified data

This is the graph between the number of tweets and number of weeks. Which indicates that in each and every week, people has done how many tweets. Such as number of +ve tweet and number of -ve tweet, number of neutral tweet.

## **CHAPTER 4:-Conclusion and Future Scope**

In today's world , amount of data is generated through communication i.e social media, organization etc. Data can be either structured form or not. To find the polarity of data initially we do sentiment analysis of data. Opinion Mining can be applied on various area such as customer feedback and marketing purpose. large number of organization are taken person feedback and perform opinion mining on these data. So that they can provide excellent services to their customer and data help the organization to upgrade their future services. Here we took some sentiment classifiers like SVM, decision tree, naive bayes which perform best in term of accuracy, recall and precision. Out of these classifier Decision Tree perform best in accuracy of twitter data. Basically our aim in this thesis is to apply opinion mining for finding the public opinion. Generally by tweets people express their feeling and opinion towards product and give review on that. so by performing sentiment analysis in these tweets finally we can able to conclude that how many person are in favour of this product and how many person are not in favour of this product so that company make their product according to customer requirement

Future scope includes, we able to make web application for this. And we can improve our classifier system such that it could handle that convey multiple sentiments (meaning). Furthermore, we can add many classification categories through that we can get better or improved results.



## Chapter 5:- References

- [1] Gurkhe D., Pal N. and Rishit B. "Effective Sentiment Analysis of Social Media Datasets using Naïve Bayesian Classification." (2014).
- [2] Bouazizi, M., Ohtsuki, T.: Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer. IEEE Access. 6, 64486-64502 (2018).
- [3] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). (2014).
- [4] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016): 1-7.
- [5] Mukherjee S., Malu A., Balamurali A.R, Bhattacharyya P. "TwiSent: A Multistage System for Analyzing Sentiment in Twitter".
- [6] Davidov D., Tsur O., Rappoport A. "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys".
- [7] Neethu, M., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). (2013).
- [8] Pulkit Garg, Himanshu Garg, VirenderRanga "Sentiment Analysis of the Uri Terror Attack Using Twitter" International Conference on Computing, Communication and Automation (ICCCA2017).
- [9] <http://www.ijcaonline.org/research/volume125/number3/dandrea-2015-ijca-905866.pdf>
- [10] <https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis>
- [11] [textblob.readthedocs.io/en/dev/modules/textblob/en/sentiments.html](https://textblob.readthedocs.io/en/dev/modules/textblob/en/sentiments.html)
- [12] <https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/>

- [13] Singla, Z., Randhawa, S., & Jain, S. (2017, July). Statistical and sentiment analysis of consumer product reviews. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
- [14] Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 437-442). IEEE.
- [15] Dinsoreanu, M., & Bacu, A. (2014, October). Unsupervised Twitter Sentiment Classification. In KMIS (pp. 220-227).
- [16] Priyadarshi, Siddhanta (23 February 2009). "Planning Commission Okays ISRO Manned Space Flight Program". *Indian Express*. p. 2.
- [17] Beary, Habib (27 January 2010). "India announces first manned space mission". Bangalore: BBC News.
- [18] E. Loper and S. Bird, "NLTK: the Natural Language Toolkit", Proc. ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics ,vol. 1,pp. 63-70, 2002.
- [19] P. Pang and L. Lee, "Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval", vol. 2(1-2), pp.1-135, 2008.
- [20] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification", Proc. AAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998
- [21] G. Kontaxis, I. Polakis, S. Ioannidis, and E.P. Markatos. Detecting social network profile cloning. In Pervasive Computing and Communications Work- shops (PERCOM Workshops), 2011 IEEE International Conference on, pages 295300. IEEE, 2011.
- [22] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* (2014).

- [23] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B.Y. Zhao. Detecting and characterizing social spam campaigns. In Proceedings of the 10th annual conference on Internet measurement, pages 3547. ACM, 2010.
- [24] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network:when bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference, pages 93102. ACM, 2011.
- [25] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schtze. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.