Project Report

On

Sentiment Analysis Of Twitter In Python Language

Submitted for partial fulfillment of requirement for the degree of

BACHELOR OF ENGINEERING

(Computer Science and Engineering)

Submitted by

Khushi Jain

Shubham Bagade

Vrushali Raut

Sharada Parachake

Under the Guidance of

Prof. S. V. Deshmukh

Department of Computer Science & Engineering,Prof. Ram Meghe Institute of Technology & Research, Badnera

2020-2021

CERTIFICATE

This is to certify that the Project (8KS07) entitled

Sentiment Analysis Of Twitter In Python Language

is a bonafide work and it is submitted to the

Sant Gadge Baba Amravati University, Amravati

By

Submitted by

Khushi Jain

Shubham Bagade

Vrushali Raut

Sharada Parachake

in the partial fulfillment of the requirement for the degree of ***Bachelor of Engineering*** *in* ***Computer Science & Engineering,*** *during the academic year* ***2020-2021*** *under my guidance.*

Prof. S. V. Deshmukh
Guide
Department of Computer Sci.& Engg.
Prof. Ram Meghe Institute Of Technology & Research, Badnera

Dr. G. R. Bamnote
Head,
Department of Computer Sci.& Engg.
Prof. Ram Meghe Institute Of Technology & Research, Badnera

External Examiner

_____

ACKNOWLEDGEMENT

With great pleasure we hereby acknowledge the help given to us by various individuals throughout the project. This Project itself is an acknowledgement to the inspiration, drive and technical assistance contributed by many individuals. This project would have never seen the light of this day without the help and guidance we have received.

We would like to express our profound thanks to Dr. G. R. Bamnote for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We would also thank the faculties of the Department of Computer Science & Engineering, for their kind co-operation and encouragement which help us in completion of this project. We owe an incalculable debt to all staffs of the Department of Computer Science & Engineering for their direct and indirect help.

Our thanks and appreciations also go to our colleague in developing the project and people who have willingly helped us out with their abilities.

We extend our heartfelt thanks to our parents, friends and well wishers for their support and timely help. Last but not the least; we thank the God Almighty for guiding us in every step of the way.

Khushi Jain
Shubham Bagade
Vrushali Raut
Sharada Parachake

Table of Contents

List of figures

Abstract

With the increase in the number of tweets in the day to day life, it is important to keep track as to which ones are correct tweet and which ones aren't. One can't judge how safe and true each tweet is based only on the comments that are mentioned for each tweet. Hence in are project we can identify the sentiment analysis of tweet that is the tweet is positive, negative or neutral. The objective is to identify the sentiment analysis of the people regarding any specific topic. So that it helps to take a future decision. Sentimental analysis is to help in determining the emotional tones behind words which are expressed in online. This method is useful in monitoring social media and helps to get a brief idea of the public's opinion on certain issues. The user cannot always get correct or true reviews about the product on the internet. We can check for user's sentimental comments on multiple applications. The reviews may be fake or genuine. Analyzing the rating and reviews together involving both user as well as admins comments. Using sentimental analysis and data mining, the machine is able to learn and Analysis the sentiments, emotions about reviews and other texts. By using sentimental analysis and data mining, analyzing reviews and comments can help to determine the correct tweet that help to take a future decision.

Keywords — Sentimental Analysis, Data mining, Review based evidence, positive negative neutral ratings, Rate evidence, Users review, Leading session.

CHAPTER 1
Introduction

1.1 INTRODUCTION TO SENTIMENT ANALYSIS

Sentiment Analysis is process of collecting and analyzing data based upon the person feelings, reviews and thoughts. Sentimental analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data.

Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then it is analyze whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment.

Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e. results of one domain cannot be applied to other domain. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing.

Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm

Fig1. Application on social media

1.2 INTRODUCTION TO PYTHON LANGUAGE

Python is a high level, dynamic programming language which is used for this thesis. Python3.4 version was used as it is a mature, versatile and robust programming language. It is an interpreted language which makes the testing and debugging extremely quickly as there is no compilation step. There are extensive open source libraries available for this version of python and a large community of users.

Python is simple yet powerful, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data, i.e. spoken English using NLTK. Other high level programming languages such as 'R' and 'Matlab' were considered because they have many benefits such as ease of use but they do not offer the same flexibility and freedom that Python can deliveressential that suspicious applications must be marked as fraud in order to be identified by the store users. It will be difficult for the user to determine the comments that they scroll past or the ratings they see is a scam or a genuine one for their benefit.

Thereby, we are proposing a system which will identify such fraudulent applications on Play or App store by providing a holistic view of ranking fraud detection system. By considering data mining and sentiment analysis, we can get a higher probability of getting real review sand hence we propose a system that intakes reviews from registered users for a single product or multiple and evaluate them as a positive or negative rating. This can also be useful to determine the fraud application and ensure mobile security.

1.3 INTRODUCTION TO NLTK

Natural Language Toolkit (NLTK) is library in Python, which provides a base for building programs and classification of data. NLTK is a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the text data in the tweets into a format that can be used to extract sentiment from them.

NLTK provides various functions which are used in pre-processing of data so that data available from twitter become fit for mining and extracting features. NLTK support various machine learning algorithms which are used for training classifier and to calculate the accuracy of different classifier.

In our thesis we use Python as our base programming language which is used for writing code snippets. NLTK is a library of Python which plays a very important role in converting natural language text to a sentiment either positive or negative. NLTK also provides different sets of data which are used for training classifiers. These datasets are structured and stored in library of NLTK, which can be accessed easily with the help of Python.

1.4 INTRODUCTION TO STREAMLIT

Streamlit is an open-sorce python library that is useful to create and share data web apps. Streamlit provides your app to stay performant even when data from the web, manipulating large datasets, or performing expensive computations. This is done with the @st. cache decorator. When you mark a function with the @st. It is an awesome new tool that allows engineers to quickly build highly interactive web applications around their data, machine learning models, and pretty much anything. The best thing about Streamlit it doesn't require any knowledge. It is an open source app framework specifically designed for ML engineers working with python . It allows you to create a stunning looking application with only a few lines of codeStreamlit is an open-source Python library that turns your scripts into shareable, interactive web applications.
Instead of writing a web application from scratch, complete with frontend interaction andbackend communication, you can simply add a couple of Streamlit functions and deploy an interactive web application in minutes.

1.5 INTRODUCTION TO WORD CLOUD

Word cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. It has been a trending technique of data visualization, especially where textual data is present. Hence, we can say that Word Cloud has been one of the prominent techniques for data visualization using Natural Language Processing (NLP).

A tag cloud(word cloud or wordle or weighted list in visual design) is a natively visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free from text.

Each word in this cloud has a variable font size and color tone . thus this representation helps to determine words of prominence, A bigger font size of a word potrays its prominence more relative to other words in the cluster. Word cloud can be built in varying shapes and sizes based on the creators vision. The number of words plays an important role while creating a word cloud. A word cloud must always be semantically meaningful and must represent what is it mean for.

1.6 INTRODUCTION TO REGULAR EXPRESSION(RE)

A Regular Expression (called Res, or regexes, or regex)is a patterns(or filter) the functions in this module let you check if a particular string matches a given regular expression ()or if a given regular expression matches a particular string, which comes down to the same thing.) Python has a module named re to work with RegEx. It is extremely useful for extracting information from text such as code, files, log, spreadsheets or even documents. RegEx in python supports various things like Modifiers, Identifiers, and White space characters.

1.7 NEED OF SENTIMENTAL ANALYSIS

1.7.1 Industry Evolution

Only the useful amount of data is required in the industry as compared to the set of complete unstructured form of the data. However the sentiment analysis done is useful for extracting the important feature from the data that will be needed solely for the purpose of industry. Sentimental Analysis will provide a great opportunity to the industries for providing value to their gain value and audience for themselves. Any of the industries with the business to consumer will get benefit from this whether it is restaurants, entertainment, hospitality, mobile customer, retail or being travel.

1.7.2 Research Demand

Another important reason that stands behind the growth of SA deals with the demand of research in evaluation, appraisals, opinion and their classification. Present solutions for the purpose of sentiment analysis and opinion mining are rapidly evolving, specifically by decreasing the amount of human effort that will be required to classify the comments. Also the research theme that will be based in the long established disciplines of computer science like as text mining, machine learning, natural language processing and artificial intelligence, voting advise applications, automated content analysis, etc**.**

1.7.3 Internet Marketing

Another important reason behind the increase in the demand of sentimental analysis is the marketing done via internet by the business and companies organization. Now they regularly monitor the opinion of the user about their brand, product, or event on blog or the social post. Thus, we see that the sentimental Analysis could also work as a tool for marketing too

1.8 APPLICATION OF SENTIMENTS ANALYSIS

Sentiment analysis has large amount of applications in the NLP domain. Due to the increase in the sentiment analysis, social network data is on high demand. Many companies have already adopted the sentimental analysis for the process of betterment. Some of major applications are mentioned as following:

1.8.1 Word of Mouth (WOM)

Word of Mouth (WOM) is the process by which the information is given from one person to another person. It would essentially help the people to take the decisions. Word of Mouth has given the information about the opinions, attitudes, reactions of consumers about the related business, services and the products or even the ones that can be shared with more than one person. Therefore, this is going to be where Sentiment Analysis comes into picture. As the online review blogs, sites, social networking sites have provided the large amount of opinions, it has helped in the process of decision-making so much easier for the user.

1.8.2 Voice of Voters

Each of the political parties usually spent a major chunk of the amount of money for the aim of campaigning for their party or for influencing the voters. Thus if the politicians know the people opinions, reviews, suggestions, these can be done with more effect. This is how process of Sentimental analysis does not only help political parties but on the other hand help the news analysts alongside. Also the British and the American administration had already used some of the similar techniques

1.8.3 Online Commerce

There is vast number of websites related to ecommerce. Majority of them had the policy of getting the feedback from its users and customers. After getting information from various area like service and quality details of the users of company users experience about features, product and any suggestions. These details and reviews have been collected by company and conversion of data into the geographical form with the updates of the recent online commerce websites who use these current techniques

1.8.4 Government

Sentiment Analysis has helped the administration for the purpose of providing various services to the public. Fair results have to be generated for analyzing the negative and positive points of government. Thus sentiment analysis is helpful in many fields like decision making policies, recruitments, taxation and evaluating social strategies. Some of the similar techniques that provide the citizen oriented government model where the services and the priorities should be provided as per the citizens. One of the interesting problems which can be taken up is applying this method in the multi- lingual country like the India where content of the generating mixture of the different languages (e.g. Bengali English) is a very common practice

CHAPTER 2

Literature Review/Survey

2.1 Literature

They were the first to work on sentiment analysis. Their main aim was to classify text by overall sentiment, not just by topic e.g., classifying movie review either positive or negative. They apply machine learning algorithm on movie review database which results that these algorithms out-perform human produced algorithms. The machine learning algorithms they use are Naïve-Bayes, maximum entropy, and support vector machines. They also conclude by examining various factors that classification of sentiment is very challenging. They show supervised machine learning algorithms are the base for sentiment analysis H. Wang, D. Can, F. Bar, S. Narayana They were the researchers who proposed a system for real time analysis of public responses for 2012 presidential elections in U.S. They collect the responses from Twitter, a micro blogging platform. Twitter is one the social network site where people share their views, thoughts and opinions on any trending topic. People responses on Twitter for election candidates in U.S. created a large amount of data, which helps to create a sentiment for each candidate and also created a prediction of whom winning.A relation is created between sentiments that arise from people response on twitter with the complete election events. They also explore how sentiment analysis affects these public events. They also show this live sentiment analysis is very fast as compared to traditional content analysis which takes many days or up to some weeks to complete. The system they demonstrated analyzes sentiment of entire Twitter data about the election, candidates, promotions, etc. and delivering results at a continuous rate. It offers media, politicians and researchers a new way which is timely effective which is completely based on public opinion.O. Almatrafi, S. Parack, B. Chavan In their research they work on Indian general elections 2014. They perform mining on 600,000 tweets which were collected over a period of 7 days for two political parties. They apply supervised machine learning approach, like Naïve-Bayes algorithm to build a classifier which can classify the tweets in either positive or negative. They identify the thoughts and opinions of users towards these two political parties in different locations and they plot their finding on India map by using a Python library. An example of their results on tweets of BJP in 2014 which shows different locations in India where BJP got positive reviews.

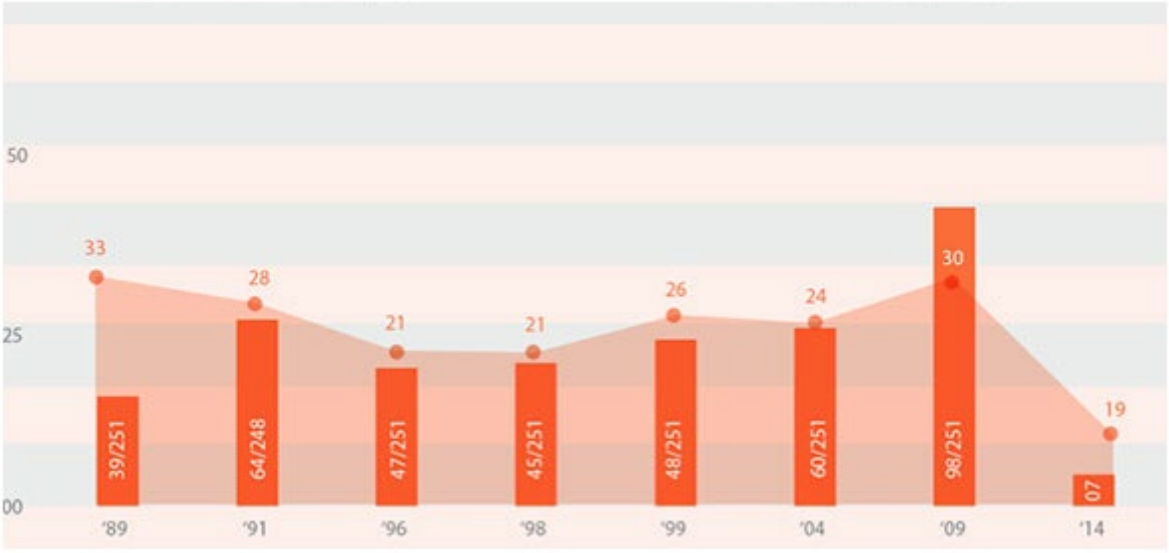


Fig 2. Tweets of BJP for different state in the year 2014

2.2 REVIEW BASED EVIDENCE

Twitter sentiment analysis was growing at faster rate as amount of data is increasing. They created a system which focuses on target dependent classification. It is based on Twitter in which a query is given first; they classify the tweets as positive, negative or neutral sentiments with respect to that query that contain sentiment as positive, negative or neutral. In their research, query sentiment serves as target. The target- independent strategy is always adopted to solve these problems with the help of state- of-the-art approaches, which may sometime assign immaterial sentiments to target. Also, when state-of-the-art approaches are used for classification they only take tweet into consideration. These approaches ignore related tweet, as they classify based on current tweet.

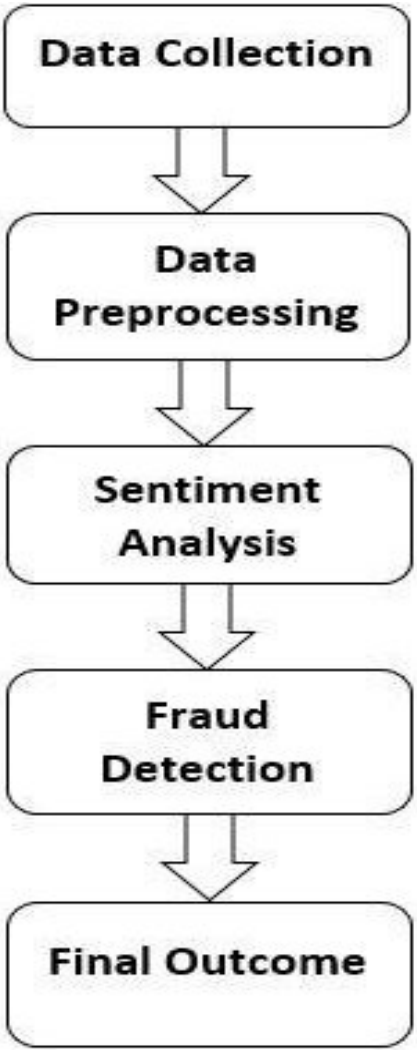


Fig3. Twitter sentiment-analysis

Data Collection

Data collection tools refer to the devices/instruments used to collect data, such as a paper questionnaire or computer-assisted interviewing system. Case Studies, Checklists, Interviews, Observation sometimes, and Surveys or Questionnaires are all tools used to collect data.

It is important to decide the tools for data collection because research is carried out in different ways and for different purposes. The objective behind data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.

Data Preprocessing

When we talk about data, we usually think of some large datasets with huge number of rows And columns. While that is a likely scenario, it is not always the case — data could be in so many different forms: Structured Tables, Images, Audio files, Videos etc.

Machines don't understand free text, image or video data as it is, they understand 1s and 0s. So probably won't be good enough if we put on a slideshow of all our images and expect our machine learning model to get trained. Data Preprocessing is that step in which the data gets transformed or encoded to bring it to such a state that now the machine can easily parse it.

Sentiment Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective I information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

Final outcome

After performing all the steps the final outcome gives us the true tweet that is fetches the most recent five tweets, generates a word cloud, perform sentiment analysis a display it in the form of a bar graph. And fetches the last hundred tweets from the twitter.

CHAPTER 3

Problem Statements

3. Problem Statement :

Sentiment Analysis is a process of extracting feature from user's thoughts, views, feelings and opinions which they post on any social network websites. The result of sentiment analysis is classification of natural language text into classes such as positive, negative and neutral. The amount of data generated from social network sites is huge; this data is unstructured and cannot give any meaningful information until it is analyzed. Thus, to make this huge amount of data useful we perform sentiment analysis, i.e. extracting feature from this data and classify them. Today, if any one wants to purchase a product or to give vote or to watch a movie, etc. then that person will first wants to know what are other people reviews, reactions and opinions about that product or candidate or movie on social media websites like Twitter, Facebook, Tumbler, etc. So there is a need of system that can automatically generate sentiment analysis from this huge amount of data. In this we can identify the tweet is positive, negative or neutral and make a decision on the basis of people tweet.

3.1 OBJECTIVE OF PROJECT

- The main objective of this thesis work is to perform the sentiment analysis on Indian Political Parties like BJP, INC and AAP

- People opinions about these parties progress, workers, policies, etc. which are extracted from Twitter.

- Thus to achieve this objective we build a classifier based on supervised learning and perform live sentiment analysis on data collected of different political parties.

3.2 Methodology

Sentiment arrangement utilizing unsupervised learning: In the unsupervised order, the content is characterized by contrasting it and given words or dictionaries. The feeling an incentive for these words or dictionaries is already characterized.

In the first step to be able to access Twitter data programmatically we need to create Sentiment and register an app on twitter developers website for authentication and thereafter we can access data by using Twitter API.

Registering App: On registering the app we will receive consumer_key and consumer secret key. Next, from the configuration page of the app, we will get access _token and access_token_secret, which will be used to get access twitter on behalf of our application. We must keep these authentication tokens private as they can be misused.

Accessing Data: Twitter provides REST API's to connect with their service. We used one python library to access twitter REST API's called Tweepy. It provides wrapper methods to easily access twitter REST API. To install Tweepy we used command pip install tweepy.

Storing Data: We access all tweet data from personal profile and store it for our analysis steps. Tweepy library provides simple cursor interface to iterate through all the tweets and store them in file.

Preparing Data: Before we begin to analyze the twitter data , it's important to understand the structure of the tweet as well as pre-process the data to remove non-useful terms means stop words. Preprocessing is in the simple term means to tka in the data and prepare the data for optimal output considering our requirement.

Data pre-processing: Pre-processing removes stop word, handling of negation; misspell correction, positive word lists of each tweet and negative word lists of each tweet.

- Filtering: Filtering is a process that removes unnecessary parts or information from
- the sentence. Filters used in many ways.
- URL: Entire URL removed from the sentence or input file after checking the whole sentence or input file. These links are replaced by the empty space.
- Usernames: Sometimes user used any username or @ symbol before any tweet. These types of usernames or @ replaced by empty space.
- Duplicate or repeated characters: Users sometimes use informal language in tweets. For example, users mostly write 'baaad' in place of the bad word.

Cleaning data: The text data here cleaned by using RegEx. Python has a built-in package called re, which can be used to work with RegEx. The re module offers a set of functions that

allows us to search a string for a match i.e used for clean the data. Functions for cleaning data we have used cleanTxt(test) function. It persform cleaning on raw text and then return the cleaned text in the form of a string and cleaning raw text but will return a list of clean words(even better). And removing digits from the text, removing the stopwords, also choose a language for applying stopwords. RegEx contain a series of text to be matched-to make a filter more specialized, or general.

- Analyze the tweet: From cleaning data first part is analyze the tweet in which fetches most five recent tweets from given tweeter handle. Analyzing of tweet is useful in generating a vast amount of sentiment data upon analysis. In this re.sub() is used replace occurences of a particular string with another sub-string. This function takes as input such as the sub-string to replace, The actual string. Suppose we wish to insert !!! instead of a white-space character in a string. This can be done via the re.sub() function as follows:
- Generate Word-cloud: After analyzing the tweet we have to clean and transform in a format that we can analyze and visualize through a word cloud. As we know RegEx is useful for replace or remove characters such as: remove the bracket, remove the extra spaces, remove punctuation Stop-words include I, he, she, and, but, was, etc. which do not add meaning to the data. So these words (tokenizing the text) must be removed which helps to reduce the features from our data. We can see the most occurring words are highlighted in bold. This helps us to understand which won't help us to understand the data better or build better model. It is a visualization technique for text that are natively used for visualizing the tags or keywords from the websites.
- Visualizing: Visualizing bigrams, which the most simultaneously occurring words. This helps us to understand which words most occur together and make our text cleaner and understanding the text distribution. Here visualization is done via plotly it ia an interactive, open source and browser-based graphing library and seaborn provides a variety of visualization patterns.

3.3 Flowchart

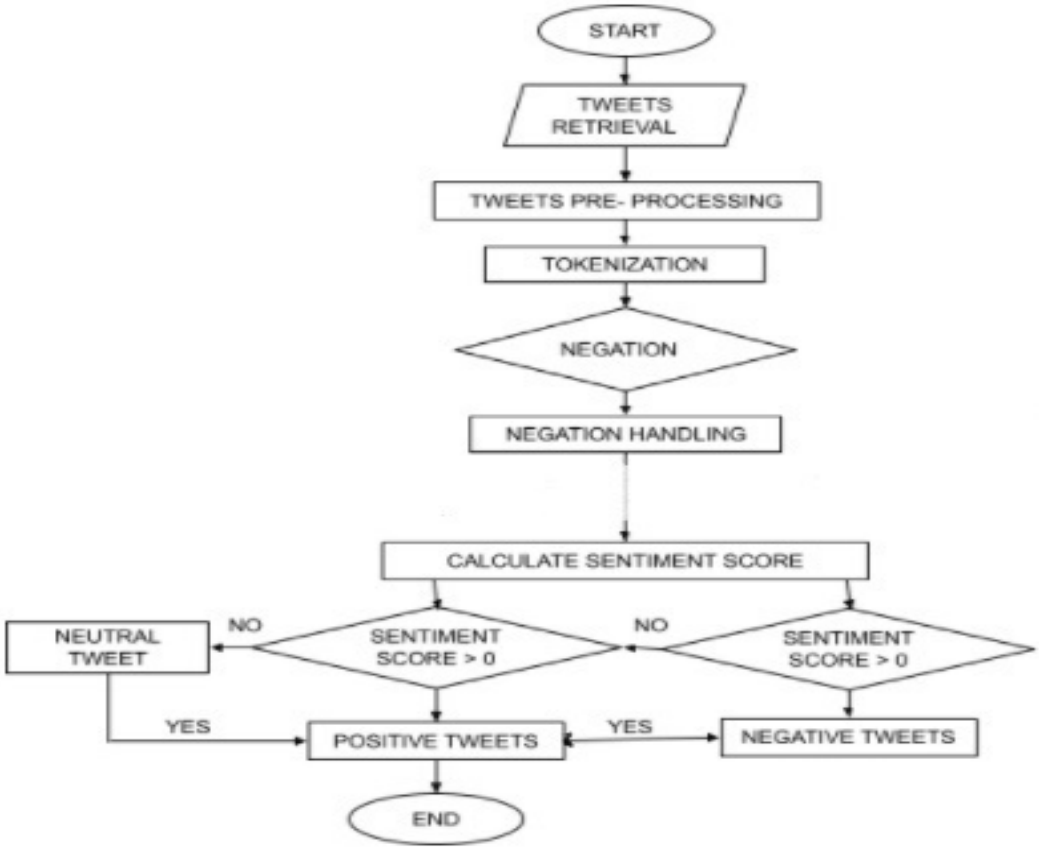


Fig 4. Sentiment Analysis of Tweet

3.4 GUI Data

Streamlit GUI:

Streamlit is an open-source Python library that can build a **UI** for various purposes, it is not limited to data apps machine learning. It is easy to learn, and a few lines of code can create a beautiful web app is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps That is:

- Make sure that you have Python 3.6 - Python 3.8 installed
- Install Streamlit using PIP : pip install streamlit

- That’s it! In the next few seconds the sample app will open in a new tab in your default browser as shown in below fig:

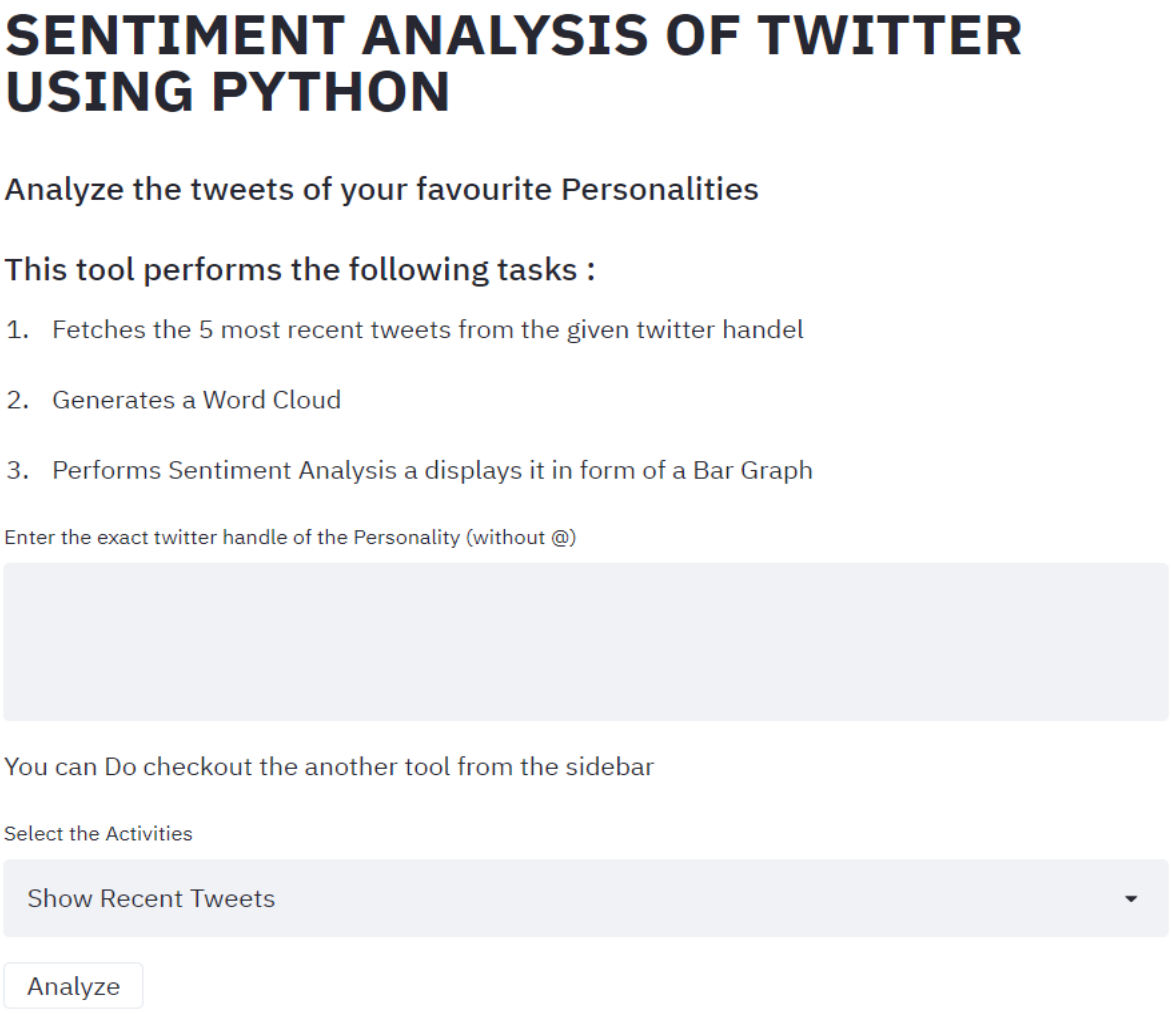


Fig 5. Streamlit GUI

CHAPTER 4

Implementation

Data collection is not a simple task, as it may seem. Various decisions have to be made for collecting data. For our thesis we maintain dataset for training, testing and for twitter sentiment analysis. In this chapter we are going to study how data is collected, stored, processed and classified. Before discussing these process and different dataset, let us discuss our proposed architecture.

4.1 Proposed Architecture:

As our goal is to achieve sentiment analysis for data provided from Twitter. We are going to build a classifier which consists of different machine learning classifiers. Once our classifier is ready and trained we are going to follow the steps shown in Figure 4.1

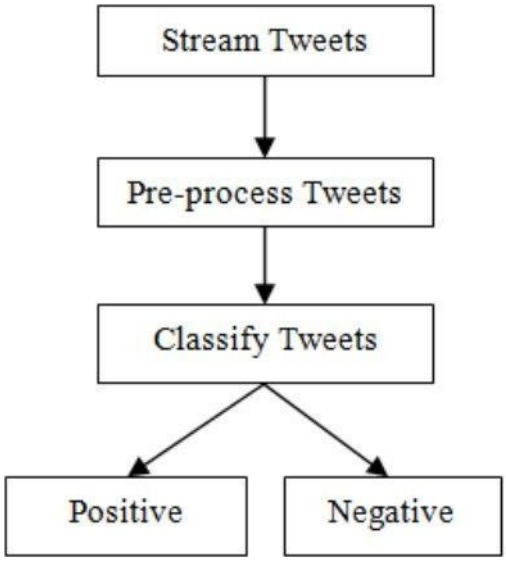


Figure 6. Process to classify tweets using build classifier

Step-1 First we are going to stream tweets in our build classifier with the help of Tweepy library in python

Step-2 Then we pre-process these tweets, so that they can be fit for mining and feature extraction.

Step-3 After pre-processing we pass this data in our trained classifier, which then classify them into positive or negative class based on trained results

4.2 Twitter API (Application Programming Interface)

Twitter allows users to collect tweets with the help of Twitter API. Twitter provides two kinds of APIs: REST API and Streaming API. The differences between these are: REST APIs support connections for short time interval and only limited data can be collected at a time, whereas Streaming API provides tweets in real-time and connection for long time. We use Streaming API for our analysis. For collecting large amount of tweets we need long-lived connection and no limit data rate

4.3 Data Collection:

Twitter Data:

To use Twitter API we must first have a twitter account. It can be easily created by filling the sign up details in twitter.com website. After this you will be provided with a username and password which is use for login purpose. Once your account is created, you can now re
send tweets on any topic you want to explore.

Twitter provider a platform from which we can access data from twitter account and can use it for our own purpose. For this we have to login with our twitter credentials in dev.twitter.com website. In this website, we first create an application which will be used for streaming tweets by providing necessary details. Once our API is created we can get to know customer key, customer secret key, access token key and access secret key. These keys are used to authenticate user when user want to access twitter data

Python is a very powerful language which provides many services with the help of many Python libraries. Tweepy is one of the open source Python library which enables Python to communicate with twitter and use its API to collect data so that we can use it in our program. To install tweepy, just provide a command 'pip install tweepy' in command prompt or bash and we ready to go with our script.

In this script we use all the keys and secrets which we got in API, we first create listener class which is used to load the data from the twitter. Now to gather data we first set up 'OAuth' protocol. OAuth is a standard protocol which is used for authorization. It allow user
any third party websites by using any social network website account without exposing passwords. OAuth provides security and authorization to user. The script which we use to

access data with the help of twitter is shown is Figure 6

```
# Create the authentication object
authenticate = tweepy.OAuthHandler(consumerKey, consumerSecret)

# Set the access token and access token secret
authenticate.set_access_token(accessToken, accessTokenSecret)

# Creating the API object while passing in auth information
api = tweepy.API(authenticate, wait_on_rate_limit = True)
```

Fig 7. Code for getting tweets using Twitter API

In this script we have to provide all the keys which are given by Twitter API. To get the tweet for a particular topic we import 'Stream' library from tweepy. In this we pass the authorization detail and the class in which we import tweets. We also apply a filter in the stream which will help us to provide the tweets for the particular topic by providing a keyword related to that topic in filter. Once we run our script, we see tweets are imported from Twitter and we can then use them for our purpose

Training Data

Other data which we collected for this thesis is training data. This data is used to train the classifier which we are going to build. To collect this data we use NLTK library of Python. NLTK consists of corpora, which is very large and consists of structured set of text files which are used to perform analysis. In these corpora there are various types of text files like quotes, reviews, chat, history, etc. From these corpora we will select files of movie reviews for our training purpose

4.4 Data Storage:

Once, we start getting our data from Twitter API our next step is to store that data so that we can use it for sentiment analysis. We ran our scripts for period of month and collect the tweets for different political parties. Every time we ran the script described in figure a .csv (comma separated values) file is generated which consists of tweets that are extracted from

Twitter API. We use .csv format for our collected data files because data consists of many fields. CSV separate each field with a comma, thus make it very easier to access the particular field which consists of text. CSV files also provide faster read/write time as compared to others

We make separate directories to store tweets of different political parties for respective month. We store them in our hard drive from where these can be easily imported to our snippet and further proceed for analysis. Once we stored our tweet we have to pre process the data stored before applying it to classifier because the data we collect from API is not fit for mining. Therefore pre-processing the data is our next step. Figure 4.3 shows glimpse of different files stored in hard drive.

Fig 8. Database of collected tweets

4.5 Data Pre-Processing:

Data obtained from twitter is not fit for extracting features. Mostly tweets consists of message along with usernames, empty spaces, special characters, stop words, emoticons, abbreviations, hash tags, time stamps, URL's ,etc. Thus to make this data fit for mining we pre-process this data by using various function of NLTK. In pre- processing we first extract our main message from the tweet, then we remove all empty spaces, stop words (like is, a, the, he, them, etc.), hash tags, repeating words, URL's, etc. We then replace all emoticons and abbreviations with their corresponding meanings like :-), =D, =), LOL, Rolf, etc. are replaced with happy or laugh. Once we are done with it, we are ready with processed tweet which is provided to classifier for required results. A sample processed tweet is shown in

Table 4.2

Table 4.2 Sample Tweet and Processed Tweet

Tweet Type	Result
Original tweet	@xyz I think Kejriwal is a habitual liar, even where he don't needs to lie he tells a lie >😒#AAP
Processed tweet	think, habit, lie, even, don't, need, tell, angry

Cleaning of Twitter data is necessary, since tweets contain several syntactic features that may not be useful for analysis. The pre-processing is done in such a way that data represented only in terms of words that can easily classify the class.

We create a code in Python in which we define a function which will be used to obtain processed tweet. This code is used to achieve the following functions:

- remove quotes - provides the user to remove quotes from the text
- remove @ - provides choice of removing the @ symbol, removing the @ along with the user name, or replace the @ and the user name with a word 'AT_USER' and add it to stop words
- remove URL (Uniform resource locator) - provides choices of removing URLs or replacing them with 'URL' word and add it to stop words
- remove RT (Re-Tweet) - removes the word RT from tweets
- remove Emoticons - remove emoticons from tweets and replace them with their specific meaning
- remove duplicates – remove all repeating words from text so that there will be no duplicates
- remove # - removes the hash tag class
- remove stop words – remove all stop words like a, he, the, and, etc which provides no meaning for classification

Table 4.3 shows the various types of contents that are included in tweets and also the actions performed on these contents. Some of the example of clean tweets is shown in Table 4.4

Table 4.3 Removed and modified content

CONTENT	ACTION
Punctuation (! ? , . ” : ;)	Removed
#word	Removed #word
@any_user	Remove @any_user or replaced with “AT_USER” and then added in stop words.
Uppercase characters	Lowercase all content
URLs and web links	Remove URLs or replaced with “URL” and then added in stop words
Number	Removed
Word not starting with alphabets	Removed
All Word	Stemmed all word (Converted into simple form)
Stop words	Removed
Emoticons	Replaced with respective meaning
White spaces	Removed

Table 4.4 Sample cleaned data

Raw data	Clean data
@jackstenhouse69 I really liked it, in my opinion it def is :)	Really, liked, opinion, def
:(\u201c@EW: How awful. Police: Driver kills 2, injures 23 at #SXSW http://t.co/8GmFiOuZbS\u201d	Sad, awful, police, driver, kills, Injures

Once our data is cleaned and ready for processing our next step

4.6 Sentiment Analysis:

Sentiment Analysis also known as Opinion mining is a relevant mining of content which recognizes and extricates emotional data in the source material and helping a business to comprehend the social slant of their image, item or administration while observing on the web discussions. Sentiment analysis is the most widely recognized content grouping device that investigations an approaching message and tells whether the basic estimation is sure, negative or unbiased.

After preprocessing of reviews system find out the sentiments of the reviews. It will classify the review as positive or negative. The system will find sentiment of the review which can be positive or negative. Positive review adds plus one to positive score, if negative it will add one to negative score. In this way it will find out score of each of the reviews and determine whether app is fraud or not on the basis of review based evidences.

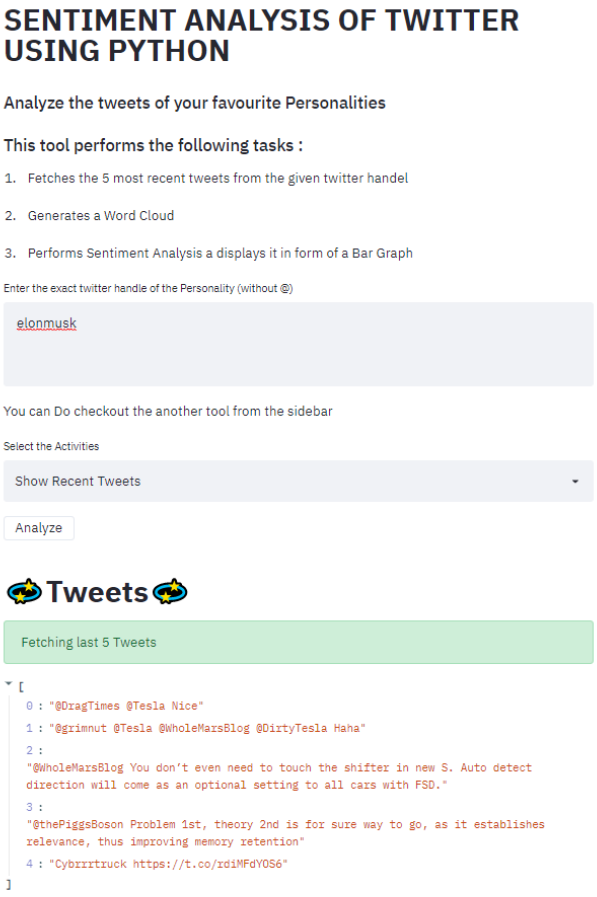
Fig 9. Sentiment Analysis of Tweet

CHAPTER 5

Analysis and Result

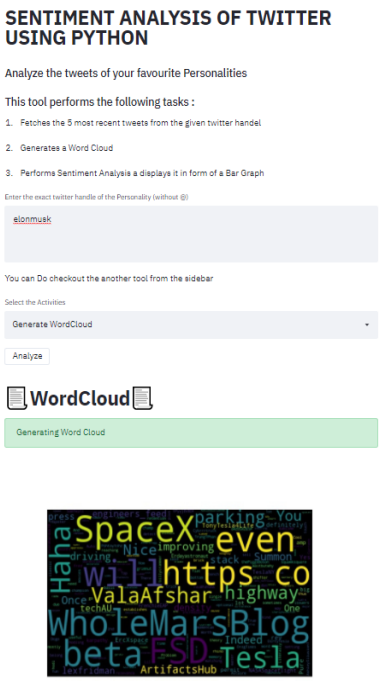
5. Tweet Collection

5.1 Fetches the Five most recent tweets from the given handel



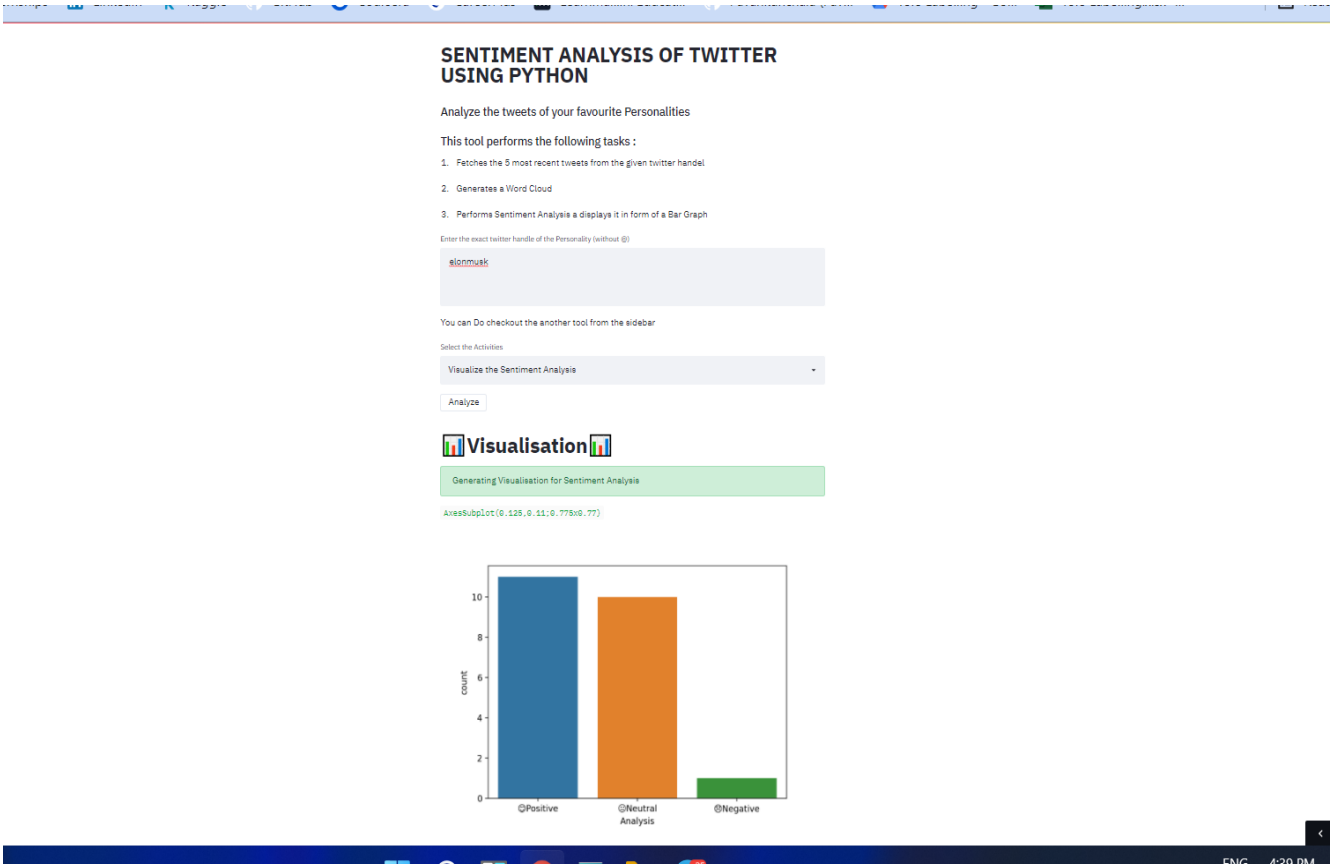
Screenshot 1. Fetches the Five recent tweet

5.2 Generates a Word Cloud



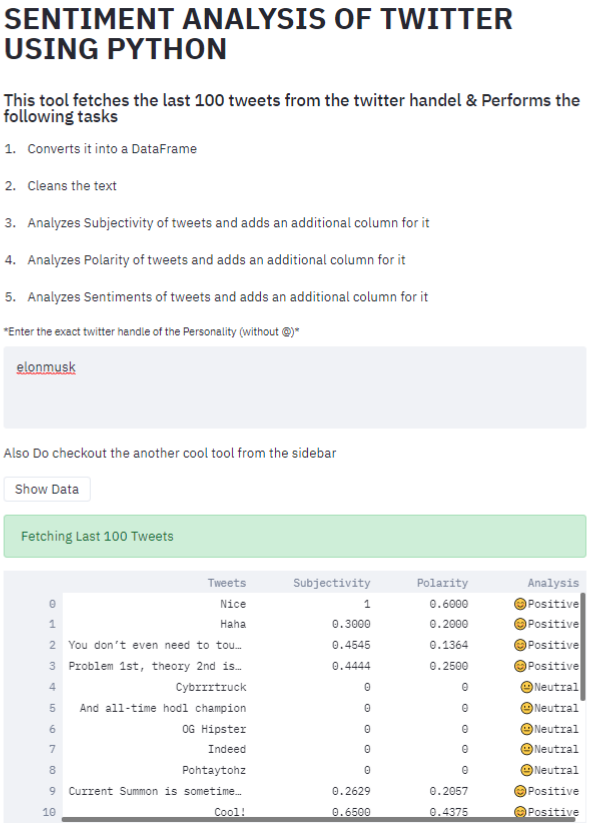
Screenshot 2. Generates a Word Cloud

5.3 Perform Sentiment Analysis and displays it in form of a Bar Graph



Screenshot 3. Perform Sentiment Analysis And Display Bar Graph

5.4 Hundred Tweets For Twitter Handel



Screenshot 4. Hundred Tweets For Twitter Handel

6. Conclusion

Sentiment analysis is used to identifying people's opinion, attitude and emotional states. The views of the people can be positive or negative. Commonly, parts of speech are used as feature to extract the sentiment of the text. An adjective plays a crucial role in identifying sentiment from parts of speech. Sometimes words having adjective and adverb are used together then it is difficult to identify sentiment and opinion. To do the sentiment analysis of tweets, the proposed system first extracts the twitter posts from twitter by user. The system can also computes the frequency of each term in tweet. We will obtain a classification of polarity of sentiments into positive, negative or neutral and prepare a plot of the same using python module like matplotlib. Twitter is large source of data, which make it more attractive for performing sentiment analysis. We perform analysis on around 100 tweets total, so that we analyze the results, understand the patterns and give a review on people opinion. We saw different people have different sentiment results according to their progress and working procedure. We also saw how any social event, speech or rally cause a fluctuation in sentiment of people. We also get to know which policies are getting more support from people which are started by any of these parties. It can be used for any purpose based on tweets we collect with the help of keyword. It can be used for finance, marketing, reviewing and many more.

7. Future Scope

Some of future scopes that can be included in our research work are:

- Use of parser can be embedded into system to improve results.
- A web-based application can be made for our work in future.
- We can improve our system that can deal with sentences of multiple meanings
- We can also increase the classification categories so that we can get better results.
- We can start work on multi language like Hindi, Spanish to provide sentiment analysis to more local.

REFERNCES

[1]H. Zang, "The optimality of Naïve-Bayes", Proc. FLAIRS, 2004

[1] C.D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval", Cambridge University Press, pp. 234-265, 2008

[2] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification", Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998

[3] M. Schmidt, N. L. Roux and F. Bach, "Minimizing finite Sums with the Stochastic Average Gradient", 2002

[4] Y. LeCun, L. Bottou, G. Orr and K. Muller, "Efficient BackProp", Proc. In Neural Networks: Tricks of the trade 1998.

[5] T. Wu, C. Lin and R. Weng, "Probality estimates for multi-class classification by pairwise coupling", Proc. JMLR-5, pp. 975-1005, 2004

[6] "Support Vector Machines" [Online], http://scikit-learn.org/stable/modules/svm.html#svm-classification, Accessed Jan 2016

[7] P. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques", Proc. ACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79-86, 2002

[8] P. Pang and L. Lee, "Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval", vol. 2(1-2), pp.1-135, 2008

[9] 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

[10] H. Wang, D. Can, F. Bar and S. Narayana, "A system for real-time Twitter sentiment analysis of 2012 U.S.presidential election cycle", Proc. ACL 2012 System Demostration, pp. 115-120, 2012

[11] O. Almatrafi, S. Parack and B. Chavan, "Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014". Proc. The 9th International Conference on Ubiquitous Information Management and Communication,2015

[12] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, "Target-dependent twitter sentiment classification", Proc. The 49th Annual Meeting of the Association

,