**A**

**MINI PROJECT REPORT**

**on**

**Twitter Sentiment Analysis**

**Submitted By**

**B.Suraj (160115737115)**

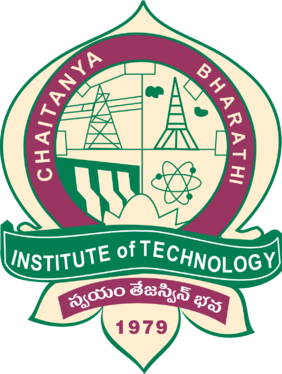
**K.Gopala Krishna (160115737113)**

Under the guidance of

# Mrs. Trupthi Mukund

**Asst. Professor,**

**Dept. of IT, CBIT.**



DEPARTMENT OF INFORMATION TECHNOLOGY

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY

**(Affiliated to Osmania University; Accreditated by NBA (AICTE), ISO Certified 9001:2000)**

**Website:** [**www.cbit.ac.in**](http://www.cbit.ac.in/)**, www.cbitworld.com**

**HYDERABAD – 500 075**

**2017**

**DECLARATION**

This is to certify that the work reported in the present report titled “**Twitter Sentiment Analysis**” is a record of work done by us in the Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad.

No part of the report is copied from books / journals / internet and wherever the portion is taken, the same has been duly referred. The reported results are based on the project work done entirely by us and not copied from any other source.

K.Gopala Krishna (160115737113)

B.Suraj (160115737115)

**ACKNOWLEDGEMENT**

It is our privilege to acknowledge with deep sense of gratitude and devotion for keen personal interest and invaluable guidance rendered by our Project Guide **Mrs. Trupthi Mukund** ,Assistant Professor, Department of Information, Chaitanya Bharathi Institute of Technology.

Our respects and regards to **Dr. Suresh Pabboju** Professor, Department of Information Technology, Chaitanya Bharathi Institute of Technology, for his invaluable suggestions that helped us in successful completion of the project.

We take the opportunity to express our thanks to **Dr.K Radhika,** Professor & Head of IT Department, CBIT for her valuable suggestions and moral support.

We are grateful to our Principal **Dr. B.Chenna Kesava Rao,** Chaitanya Bharathi Institute of Technology, for his cooperation and encouragement.

Finally, we also thank all the staff members, faculty of Dept. of IT, CBIT, our friends, and all our family members who with their valuable suggestions and support, directly or indirectly helped us in completing this project work.

K.Gopala Krishna(160115737113)

B.Suraj(160115737115)

**ABSTRACT**

[Twitter](http://www.twitter.com/) is a popular social network where users can share short SMS-like messages called tweets. Users share thoughts, links and pictures on Twitter, journalists comment on live events, companies promote products and engage with customers. The list of different ways to use Twitter could be really long, and with 500 millions of tweets per day. This gives a lot of data to mine and analyze.

This Mini Project acts as a data science analysis tool which provides the twitter users to differentiate and analyze the tweets and come to conclusions. With help of twitter and a good data science tool, you can find out what the world is saying about a particular topic. Be it reviews about movies, support to teams in games, sentiments about elections or any hot topic off the press – you can know what the people are saying by yourself.

[Sentiment Analysis](http://en.wikipedia.org/wiki/Sentiment_analysis) is one of the interesting applications of data analytics. Although the term is often associated with [sentiment classification of documents](https://marcobonzanini.com/2015/01/19/sentiment-analysis-with-python-and-scikit-learn/), broadly speaking it refers to the use of text analytics approaches applied to the set of problems related to identifying and extracting subjective material in text sources. This project can be used in social media analytics where sentiment analysis can evaluated. Performing this exercise helps you understand some of the challenges in mining social media.

Positive tweets and Negative tweets can be identified using this tool, fans support for a team in sports or support for a movie can be easily identified. Graphs and tabular columns are used to represent the analyzed data.

**Technologies and Tools :**

* Python
* NLTK – Natural Language Tool Kit
* **Numpy**
* **Pandas**
* **Matplotlib.**

**Introduction**

* 1. **Overview**

This report discusses the result of the work done in the formulation of this program of “Twitter Sentiment Analysis” on the Python platform.

It is the part of mini project lab course going on in Information Technology Department, Chaitanya Bharathi Institute of Technology.

During the course of the project we had to develop a program that extracts tweets from twitter on a certain topic and these extracted tweets are classified in to either positive or negative and a respective graph is plotted accordingly.

**1.2 Advantages of this Project:**

* It’s a user friendly project where even naïve users can use it effectively
* It helps the user to know what people are thinking around the world.
* It is helpful to predict future trends.
* **Helps in market based analysis.**
* Voice of Voters
* Brand Reputation Management (BRM)
  1. **Aim of the Project**

With the emergence of social networking, many websites have evolved in the past decade like Twitter, Facebook, Tumbler, etc. Twitter is one the website which is widely used all over the world. According to Twitter it has been recorded that around 200 billion tweets posts every year. Twitter allows people to express their thoughts, feelings, emotions, opinions, reviews, etc. about any topic in natural language within 140 characters. Python is the standard high-level programming language which is best for NLP. Thus, for processing natural language data, Python uses one of its libraries called Natural Language Toolkit (NLTK). NLTK provides large amount of corpora which helps in training classifiers and it helps in performing all NLP methodology like tokenizing, part-of-speech tagging, stemming, lemmatizing, parsing and performing sentiment analysis for given datasets.

It is a challenging task to deal with a large dataset, but with the use of NLTK we can easily classify our data and give more accurate results based on different classifiers. The goal of this thesis is to perform sentiment analysis on different Indian Political Parties, like BJP (Bharatiya Janta Party), AAP (Aam Aadmi Party), INC (Indian National Congress). Public opinions of these parties are mined from Twitter and then classified into sentiments, whether positive or negative by using supervised machine learning classifiers. These results will let us know about the reviews and opinions of people on these political parties.

To achieve this goal, a module is created which can perform live sentimental analysis. In live sentimental analysis user can obtain the trend of any live trending topic depicted by two sentiment category (positive and negative) in live graphs. Further accuracy and reliability of the module can be checked with the help of various machine learning classifiers.

To many companies and organizations a customer’s perception of a product or service is extremely valuable information. From the knowledge gained from an analysis such as this a company can identify issues with their products, spot trends before their competitors, create improved communications with their target audience, and gain valuable insight into how effective their marketing campaigns were. Through this 8 knowledge companies gain valuable feedback which allows them to further develop the next generation of their product.

In this project we work on different political parties because in our country politics plays a very vital role. Winning an election by any party is different from how that party works after winning.

In the context of the sentiment analysis being carried out for this application, the results will allow user to gain insight into how each party is being perceived by the public. This is very valuable information as public is uploading their expectations, opinions and views on the political parties. This really revolutionizes the feedback process. An application such as this has the potential to analyze the sentiment in real time giving the users immediate feedback on how a party is being help in the eyes of its audience. Such an application could be expanded to use clustering algorithms to give insight into particular member or position.

* 1. **Organization of the Report**

**Sentiment Analysis in Python**

**2.1 Introduction to Python**

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 deliver.

**2.2 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 1 sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm.

**2.3 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 project 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.

**Software Requirement Specification**

**3.1 Introduction**

The requirements specification is a technical specification of requirements for the software products. It is the first step in the requirements analysis process it lists the requirements of a particular software system including functional, performance and security requirements. The requirements also provide usage scenarios from a user, an operational and an administrative perspective. The purpose of software requirements specification is to provide a detailed overview of the software project, its parameters and goals. This describes the project target audience and its user interface, hardware and software requirements. It defines how the client, team and audience see the project and its functionality.

**3.2 Purpose of the document:**

The purpose of this document to inform the reader. It is helpful, both to the reader and to the writer, if the report is logically organized.

This document is an in paper version of the ideas and implementation of this project.

It contains the Aim, software requirements, implementation aspects, algorithms and flowcharts.

**3.3 Software Requirement Specification**

|  |  |
| --- | --- |
| Operating System | Windows XP/7/8.1/10 |
| Programming Languages | Python 2.7.6 or more. |
| Processor | Intel(R) Core(TM) i3/i5/i7 CPU M 350 @2.27GHz |
| RAM | 2 GB or more |
| Disk Space | 4GB or more |
| Modules required to be installed in Command Prompt | Scipy, Scikit-learn, Numpy, Pandas, NLTK, Tweepy | |

**Implementation**

**4.1 Introduction**

The success of the software product is determined only when it is successfully implemented according to the requirements. The analysis and the design of the proposed system provide a perfect platform to implement the idea using the specified technology in the desired environment. The implementation of our system is made user friendly.

Any software project is designed in modules and the project is said to be successfully implemented when each of the module is executed individually to obtain the expected result and also, when all the modules are integrated and run together without any errors.

**4.2 Mechanism of the Program:**

**4.2.1 Flowchart:-**

Stream Tweets

Pre-Process Tweets

Negative

Tweets

Positive

Tweets

Classify Tweets

**4.3 Methodology**

**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.

**4.3.1 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.

**4.3.2 Data Collection or Data Streaming :**

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 read and 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. As the objective of this project is to analyze the sentiment of Tweets posed for political parties, only tweets about related to this should be collected. Hence for this we create a Python script which will be used to fetch tweets from twitter. Before creating this script we first install a library in Python called **tweepy.**

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 to log in any third party websites by using any social network website account without exposing passwords. OAuth provides security and authorization to user.

**4.3.3 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.

**4.3.4 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.

**4.3.5 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 preprocessing 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.

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

Original Tweet

Pre-Processing Tweet

think, habit, lie, even, don’t, need, tell, angry

Pre-Processed

(An example of pre processed tweet)

**4.3.6 Classification**

To classify tweets in different class (positive and negative) we build a classifier which consists of several machine learning classifiers. To build our classifier we used a library of Python called, Scikit-learn. Scikit-learn is a very powerful and most useful library in Python which provides many classification algorithms. Scikit-learn also include tools for classification, clustering, regression and visualization. To install Scikit-learn we simply use on line command in python which is ‘pip install scikitlearn’.

In order to build our classifier, we use Naïve-Bayes Classifier.

**Testing and Results**

**5.1 Introduction**

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. In fact, testing is the one step in the software engineering process that could be viewed as destructive rather than constructive.

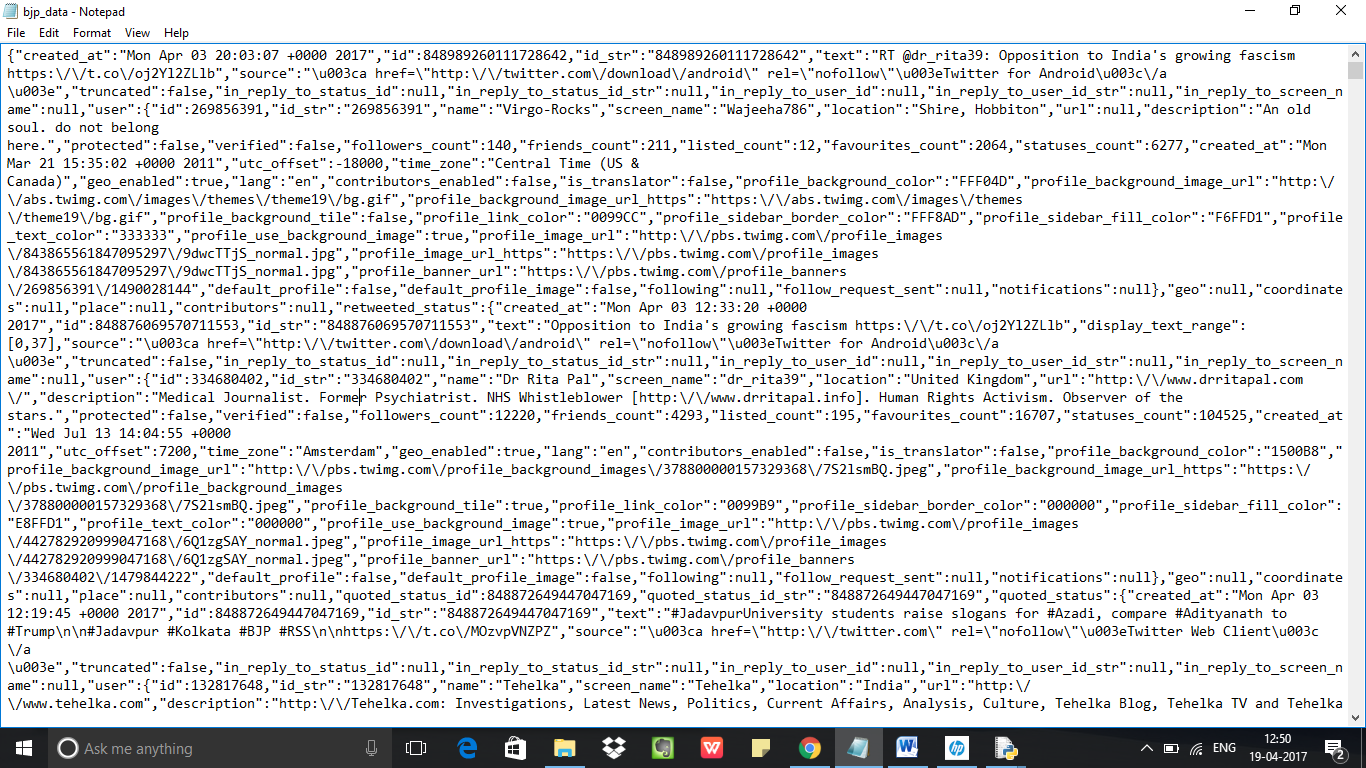
A strategy for software testing integrates software test case design methods into a well-planned series of steps that result in the successful construction of software. Testing is the set of activities that can be planned in advance and conducted systematically. The underlying motivation of program testing is to affirm software quality with methods that can economically and effectively apply to both strategic to both large and small-scale systems.

**5.2 Testing Objectives**

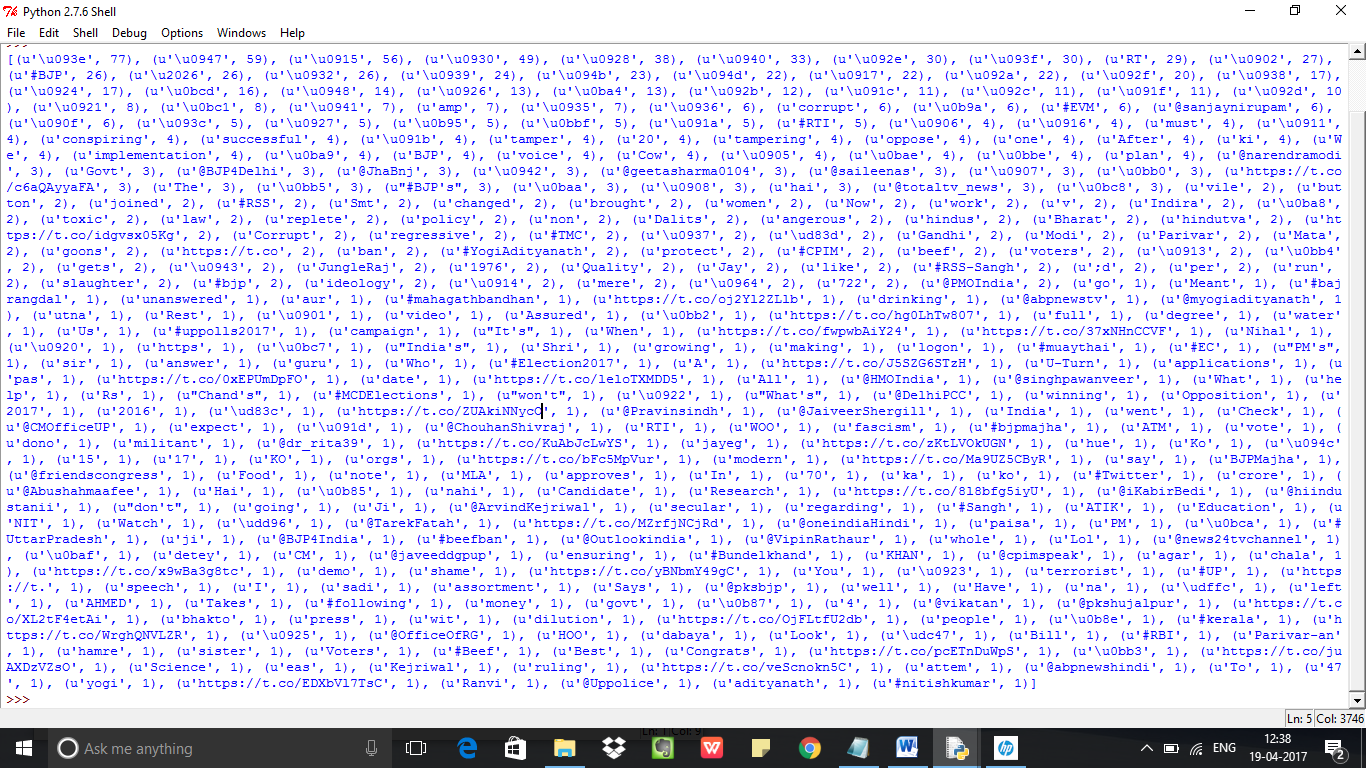
The main objective of  performance testing is designed to test the runtime performance of the system within context of the systemand whether the timetables of respective staff and class given by the user are displayed or not.These tests were performed at module level as well as system level. Individual modules were tested for required performance .

As the test results are gathered and evaluated they begin to give a qualitative indication of the reliability of the software. If proper output is not obtained, the overall quality of the software is a natural suspect. If, on the other hand, all the results which are not successful, are encountered, and are easily modifiable, then the following conclusion can be made: The tests are inadequate as the requirements mentioned are not compatible. The testing includes:

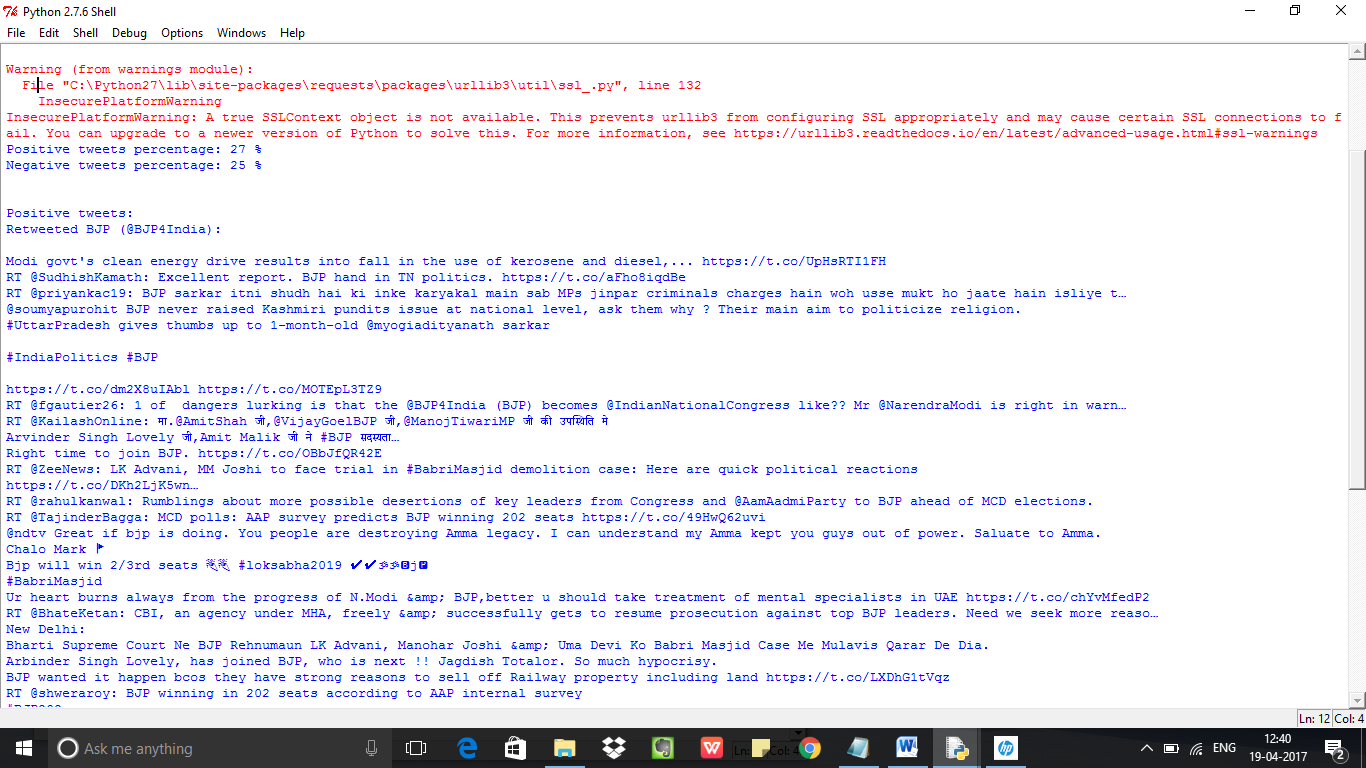
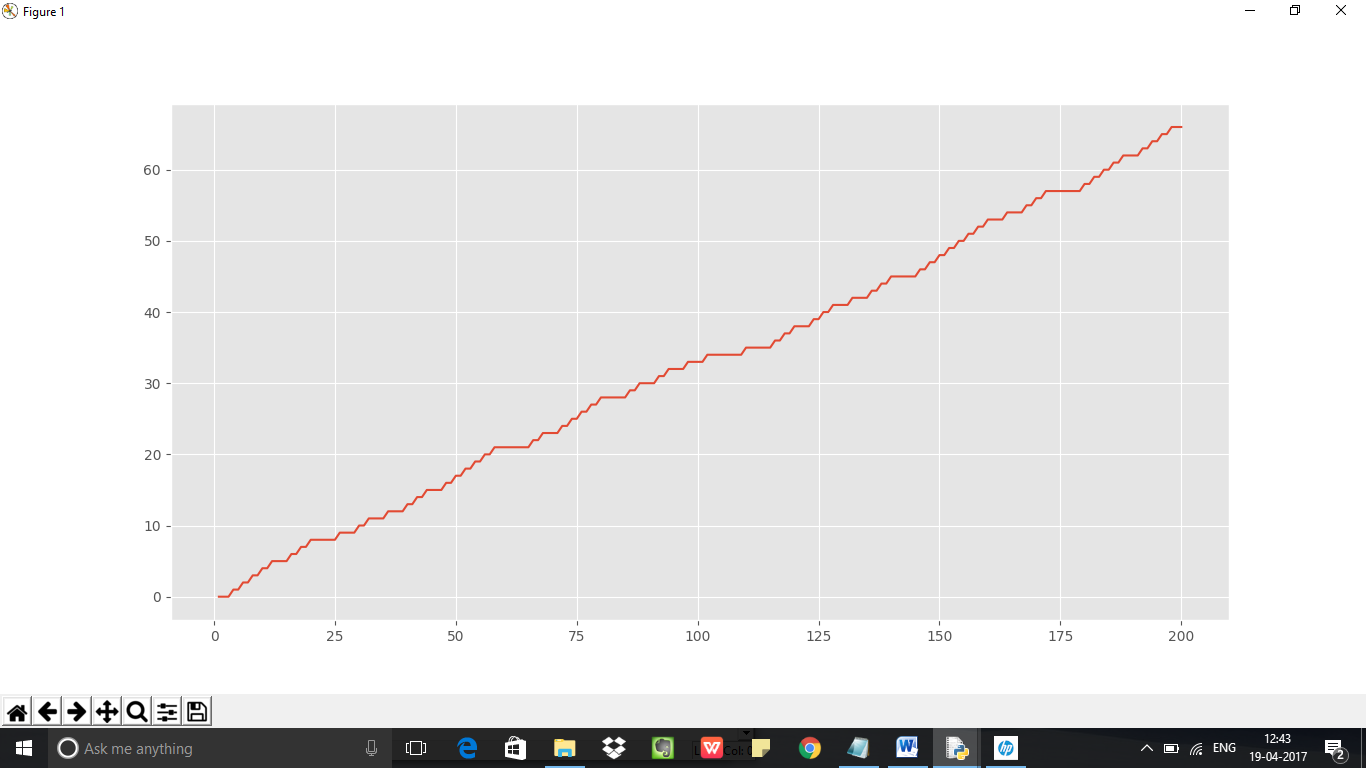
* The position and related labels for all controls were checked.
* Name of the form in system is given appropriately.
* All menu functions and sub functions were verified for correctness.
* Validations for all inputs were done.
* Each menu functions were tested, whether it invokes the corresponding functionality properly
* Whether the non-editable text control is disabling and it was also verified that it doesn't exceed the maximum allowed length. Whether the system prompts the user with appropriate message as and when invalid information is entered. All required fields aren't left blank.

**5.3 Output Screens**

**Figure 5.3.1 Streamed Data.**

****

**Figure 5.3.2 Pre-Processed data.**

**Figure 5.3.3 Positive Percentage of tweets and Negative Percentage of Tweets.**

**Figure 5.3.4 Graph for streamed data. This indicates that the sentiment of BJP is good( As there are very few negative and neutral tweets.)**

**Conclusion and Future Scope**

**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. Using machine learning supervised approach help to obtain the results. Twitter is large source of data, which make it more attractive for performing sentiment analysis. We perform analysis on around 15,000 tweets total for each party, so that we analyze the results, understand the patterns and give a review on people opinion. We saw different party 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 was shown that BJP is more successful political part in present time based on people opinion. It is not necessary that our classifier can only be used for political parties. It is general classifier. 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.

**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 languages like Hindi, Spanish, and Arabic to provide sentiment analysis to more local.

**Bibliography**

The following websites has helped us to understand the concepts of Java and XML in a very user friendly way.

* <https://marcobonzanini.com/>
* <https://pythonprogramming.net/twitter-sentiment-analysis-nltk-tutorial/>