**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

**Sentimental Analysis of Twitter Data**

Submitted by

Krunal Bhatt 1MS12CS048

Nishant Washisth 1MS12CS069

Raghav Chawla 1MS12CS079

Rajat Jain 1MS12CS085

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

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), **May 2015**

**Acknowledgements**

First and foremost,my utmost Gratitude to our HOD, Dr. K G Srinivasa who gave us a chance to complete this project. He has been our inspiration as we overcome all the obstacles in the completion of this project work. He had kind concern and consideration regarding project work which helped us to successfully complete the project.

Secondly we would like to thank Chandrika Prasad, Assistant Professor,MSRIT who reviewed our project and gave us a valuable guidance which helped us in completing this project.

This work would not have been possible without the guidance and help of several individuals who in one way or another contributed their valuable assistance in preparation and completion of this study.

We would like to express sincere thanks to all the teaching and non-teaching faculty of CSE department and my dear friends who helped in all the ways while preparing the Report.

# Abstract

Twitter, one of the largest social media site receives tweets in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This paper provides a way of analysing twitter data.

The main purpose of this project is to collect large data from social networking site 'TWITTER' and analyse this data which can depict the public sentiments and emotions that tells the nature of public .The aim of this project is to come up with a good result that helps the public from the analysis of user tweets. This includes identifying the locations where the tweets are made more and differentiating between positive, negative and neutral tweets.

#### Contents

***Declaration i***

***Acknowledgements ii***

***Abstract iii***

1 **INTRODUCTION**

* 1. General Introduction
  2. Statement of the Problem
  3. Objectives of the project
  4. Project deliverables
  5. Current Scope
  6. Future Scope

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1 Introduction

3.2 Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.3 Functional Requirements

4.3.1 Functional Requirement 1.1

:

4.3.n Functional Requirement 1.n

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.5 Performance Requirements

4.6 Database Requirement

4.7 Design Constraints

4.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools used
   2. Technology used
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**

**7.1** Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK**
2. **REFERENCES**
3. **INTRODUCTION**
   1. General Introduction

The rise of micro blogging services like Twitter has spawned great interest in these systems as human-powered sensing networks. Since its creation in 2006, Twitter has experienced an exponential explosion in its user base, reaching a lot of people across the globe. So analysis on the twitter data will help the people in knowing the mindset of the people around us.

Hence in this project we are doing an overall analysis of the twitter dataset which will classify the tweets as positive, negative or neutral. The main technologies used in this are eclipse, Hadoop and map reduce parallel programming concept. The detailed design and implementation is explained below. Here we concentrate on the analysing the dataset and visualising the output using pie chart which is done using python.

* 1. Problem Statement

The purpose of this project is to find out trends by aggregating the data in social networking site such as Twitter. Analysis of Twitter data to evaluate sentiment of the user tweets and classify them as positive, negative or neutral tweets.

1.3 Objectives of the Project

The objective is to implement an algorithm for automatic classification of text into positive, negative or neutral. Sentiment Analysis to determine the attitude of the mass is positive, negative or neutral towards the subject of interest and Graphical representation of the sentiment in form of Pie-Chart.

1.4 Project Deliverables

The PPT has been prepared and the document has been made on this topic and same thing has been presented which explains the above mentioned things.

The SRS has been prepared and the Design document has been made on this topic and same thing has been presented which explains the above mentioned things.

The output is being visualized using Pie-chart by classifying the tweets as positive, negative and neutral.

1.5 Current Scope

Twitter has lot of scope in the modern era. Twitter as a social media has many users and the number of users is increasing day by day. At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot ’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret.

1.6 Future Scope

Nowadays big data has become the buzzword in IT industry organizations. The need of analysing and processing of information has grown a lot. This paper implemented the analysing of big data (tweets) only for text. Further analysis can be done to images and all types of multimedia files based on index support. The result of Text mining and data analysis would help in suggesting related pages based on different types of data. So that industries make the data easily available to people who is using and trying accessing such type of data.

1. **PROJECT ORGANIZATION**

2.1 Software Process Models

**Agile Model**: We choose to follow it based on the project size, complexity and duration. The agile methodology is a defined process that makes it easier for a project team to deal with complexity and coherence using an incremental delivery approach. If project requirements are clearly defined and understood and expected project duration is 3 to 4 months and hence we suggest iterative and incremental development model with periodic reviews.



This model reduces the project development risk because of multiple incremental elaboration, construction cycle and addition of new features. In addition to four basic phases such as inception, elaboration, construction and transition, we provide support for project specific timeframe.

2.2 Roles and Responsibilities

The roles and responsibilities for the project are-

* Rajat Jain- Maintaining the backend and linking the functionality to the frontend. Technologies used are twitter streaming API.
* Raghav Chawla- Normalizing the dataset, stemming and implementing Naïve Bayes algorithm. Technologies used are NLTK library for python, Naïve Bayes algorithm and Python.
* Nishant Washisth- Collecting the dataset, performing the tests, documenting the details. Technologies used are twitter streaming API.
* Krunal Bhatt- Classifying the dataset as positive, negative and neutral using python and monitoring the project. Technologies used are Python and NLTK library for Python.

1. **LITERATURE SURVEY**

3.1 Introduction

The rise of micro blogging services like Twitter has spawned great interest in these systems as human-powered sensing networks. Since its creation in 2006, Twitter has experienced an exponential explosion in its user base, reaching a lot of people across the globe. So analysis on the twitter data will help the people in knowing the mindset of the people around us.

Hence in this project we are doing an overall analysis of the twitter dataset which will classify the tweets as positive, negative or neutral. The main technologies used in this are Twitter streaming API, NLTK library for python and Naive Bayes algorithm . The detailed design and implementation is explained below. Here we concentrate on the analysing the dataset and visualising the output using pie chart which is done using python.

3.2 Main Body

1. **Introduction**

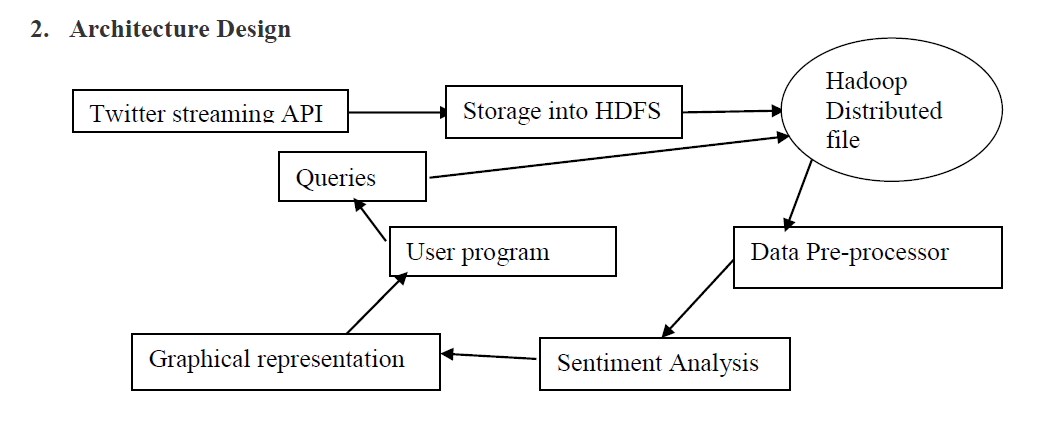
a. Number of Modules – 3

b. Modules description

Training set contains a set of pre-determined positive and negative tweets. The training set is used to identify words with features.

Feature extractor can apply the features to our classifier. We pass the feature extractor along with the tweets list defined above.

The Naive Bayes classifier uses the prior probability of each label which is the frequency of each label in the training set, and the contribution from each feature.



You get the probability for a text to belong to each of the categories you test against. The category with the highest probability for the given text wins:

In our case the probability of positive or negative is calculated.

3.3 Conclusion of Survey

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to java twitter streaming API. It was very interesting to gather and then aggregate the social networking data so as to extract interesting patterns and recent trends from it. We got exposure to work with prominent parallel data processing tool: Twitter streaming API, NLTK library for python, Naïve Bayes Algorithm and Python Language is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly.

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

1.External Interface Requirements

This section contains all of the functional and quality requirements of the system. It gives a detailed description of the system and all its features.

a) User Interfaces:

The application would provide 3 main user interfaces.

The first screen would allow the user to enter their search query parameters. It would include text entry fields and submit buttons.

The user can choose between basic search would allow them to enter a query word (hash tag) while advanced search would allow the user to enter multiple words that should /should not be present in the tweet.

Upon submitting the search the user will see a loading screen which shows the progress of the application. The progress steps which the user would see are:

***Gathering data - Tokenizing sentence - Feature extraction – Classification - Generating analytics***

The analytics screen would show the user the graphical distribution of the results. The identified tweets would be calculated and distributed by positive and negative sentiment of tweets according to location.

b) Hardware Interfaces

There are no significant hardware interfaces besides basic input and output required by the application.

c) Software Interfaces

The Twitter streamer API grabs tweets and stores them after pre-processing the data.

When the user launches the application and searches for a hashtag the Naïve Bayes algorithm searches the data and retrieves it.

The data then goes through the classifying algorithm and the result is generated and sentiment is identified.

d) Communication Interfaces

Communication between all the software interfaces is needed. This communication is usually triggered by the user application.

2. Functional requirements

User class 1

Functional Requirement 1.1

Feature: Basic search

When the user enters a hashtag to be searched and presses the search button the program will launch the sentiment analysis program. It will search for all tweets in the database that contain the hashtag the user has entered.

The application will identify and inform the user if the hashtag in invalid or not found in the database.

Functional Requirement 1.2

Feature: Advanced search

The user can select the option to advance search. This will reveal more options.

The user has the added option to search for tweets witch contain the words entered or do not contain the words entered. The user may enter some words separated by spaces in both the fields.

When the search button is pressed the application will search for tweets with the given hashtag and all include all the word entered in the **include these words** field and not include the words in the **do not include these words** field only.

Functional Requirement 1.3

Feature: Interactive graphical representation.

Once the results are returned the graphical representation of the results will be displayed. It will contain the percentage of identified positive and negative tweets on the search query. The data can be sorted by location and time. This feature would allow the application to be more interactive and informative to the user.

3. Software System Attributes

a) Reliability

The application should get at least 1000 tweets for any given search query from a random pick of tweets in database for proper results. The ideal value is 100% of the time however realistic value is 99% occurrence.

The network connection should be reliable when data retrieval is being done.

b) Availability

The availability of the system when it is used should be maximum. The ideal value is 100% not considering network failure.

c) Security

The system storage should be secure from threats and malicious programs. They could harm the data and alter it which would affect the results. The cloud storage should be secure for the same reason

d) Portability

The application is written in python and can be run on any system. The data retrieval will be done from a cloud storage.

e) Maintainability

The application should be easy to extend. The code should be written in a way that it favours implementation of new functions.

Test environments should be built for the application to allow testing of the applications different functions.

f) Performance Requirements

The application should be able to retrieve the results in the fastest time possible. The code must be optimized in order to increase the performance.

g) Database Requirement

The cloud database should be able to hold the required number of entries (1,00,000).

There should be a temporary database on the system which would hold the data related to the search query and another to hold the final result to be displayed.

h) Design Constraints

This section includes the design constraints on the software caused by the hardware.

The application needs sufficient hard drive space to be stored. The program cannot occupy too much hard drive space.

The application needs sufficient memory to run on the system. The program cannot exceed the available memory on the system.

i) Other Requirements

The program requires a database of word with their classification based on sentiment. The database is a test set to train the program to identify the sentiment in the tweets

1. **DESIGN**

1. **Introduction**

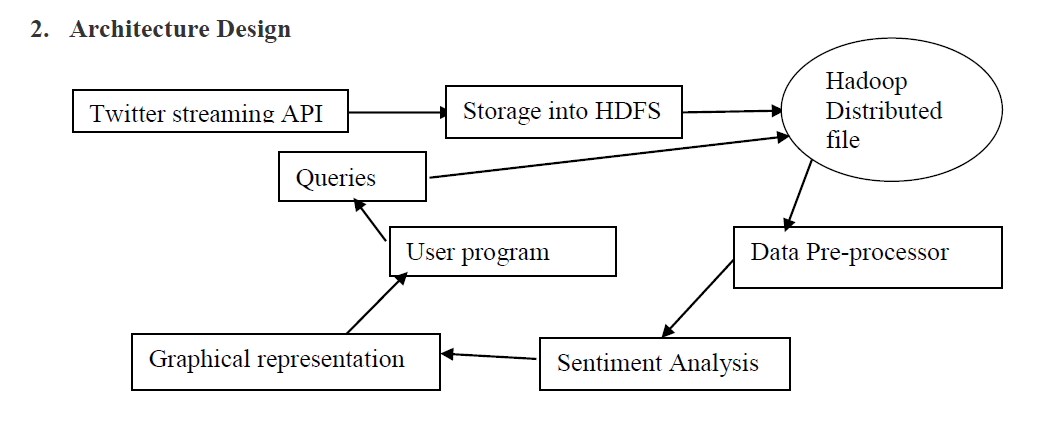
a. Number of Modules – 3

b. Modules description

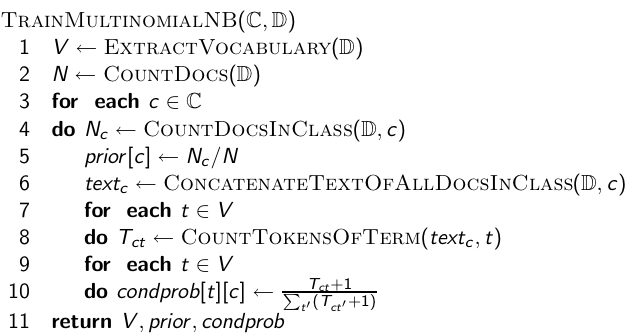
Training set contains a set of pre-determined positive and negative tweets. The training set is used to identify words with features.

Feature extractor can apply the features to our classifier. We pass the feature extractor along with the tweets list defined above.

The Naive Bayes classifier uses the prior probability of each label which is the frequency of each label in the training set, and the contribution from each feature.



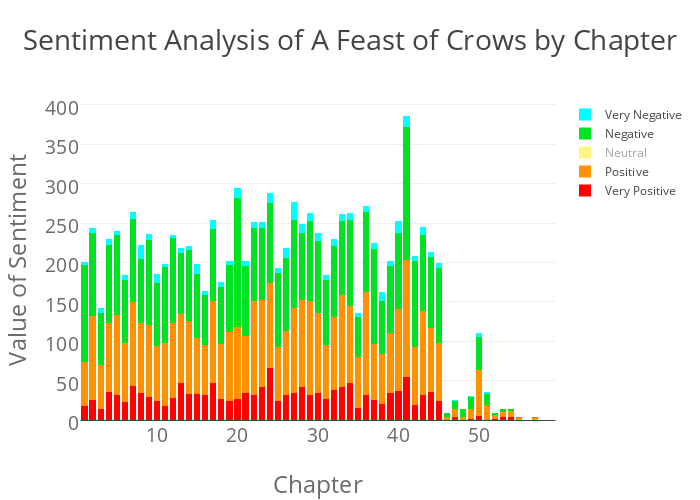
**Naive bayes algorithm (classification)**

******

**Example: **

Graphical User Interface

The interface will meet the following requirements to conform to the users’ needs. It will be simple and easy to understand. Controls which allow the user to interact with the application will be clear and imply their functionality within the application. The user interface of eclipse was simple to navigate and easy to understand.

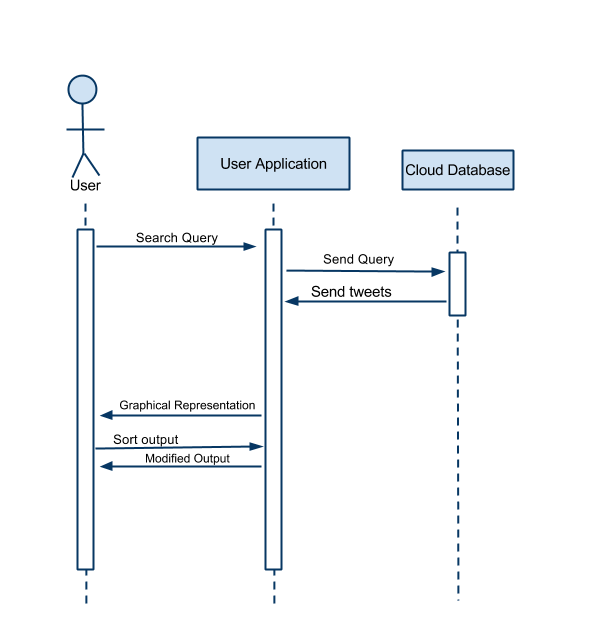


The output is a real time data in the form of a simple graph. The graph will illustrate the nature of tweets over different geolocations. This output should be clear and easy to understand.

Class Diagram/ Data flow diagram:



Sequence Diagram:

****

**6. IMPLEMENTATION**

6.1 Tools Used:

Twitter Streaming API, NLTK Library for Python.

6.2 Technology Used:

NLTK library, Naive Bayes Algorithm, Python.

6.3 Overall view of the project in terms of implementation

The Project is implemented as follows:

* Twitter Data Set Identification and Extraction
* Pre-processing of data set
  + Removing unwanted attributes
  + Eliminating stop words
  + Stemming
* Tweet classification
  + Positive
  + Negative
  + Neutral

Explanation of Algorithm and how it is been implemented

**Our Approach In Implementation**:

In our approach we focused more on the speed of performing analysis than its accuracy i.e. performing sentiment analysis on big data which is achieved by splitting the various modules of data in following steps and collaborating with hadoop for mapping it onto different machines. Part of speech tagged using opennlp. This tagging is used for following various purposes.

i. **Stop words removal:**

The stop words like a, an, this which are not useful in performing the analysis are removed in this phase. Stop words are removed using java in eclipse platform. All the words are not considered are not considered for analysis.

ii. **Unstructured to structured**:

Twitter comments are mostly unstructured i.e. ‘aswm’ is written ‘awesome’, ‘happyyyyyy’ to actually ‘happy’. Conversion to structured is done by dynamic data records of unstructured to structured and vowels adding.

iii. **Stemming**:

In stemming we remove suffix and prefixes from the tweeter data set file and normalise it.

.

6.5 Information about the implementation of Modules

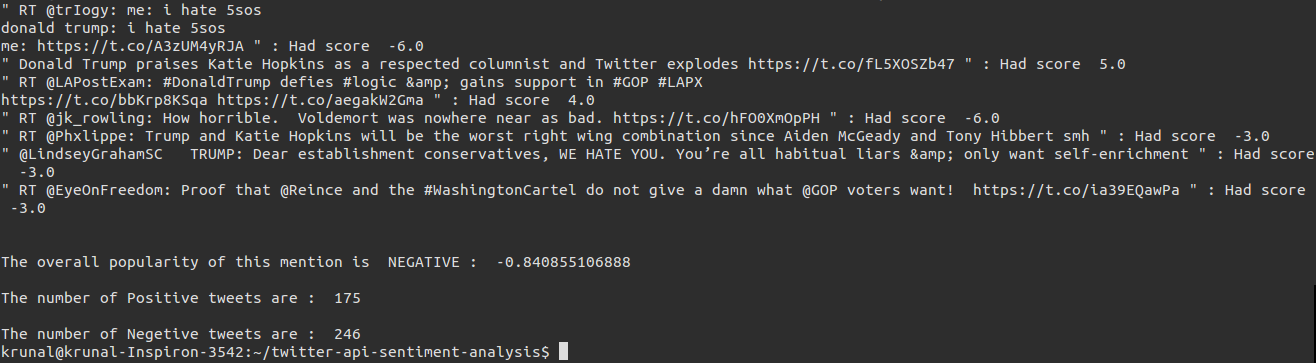
Number of Modules: 4

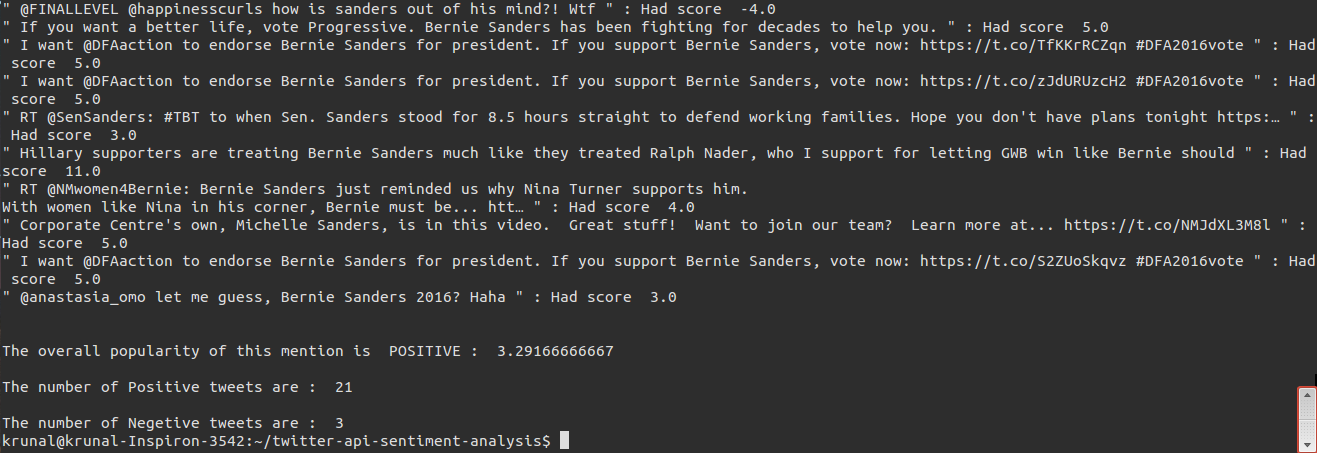
* Fetching and Extracting Data:
* Data Pre-processing:
* Classification:
* Analysis

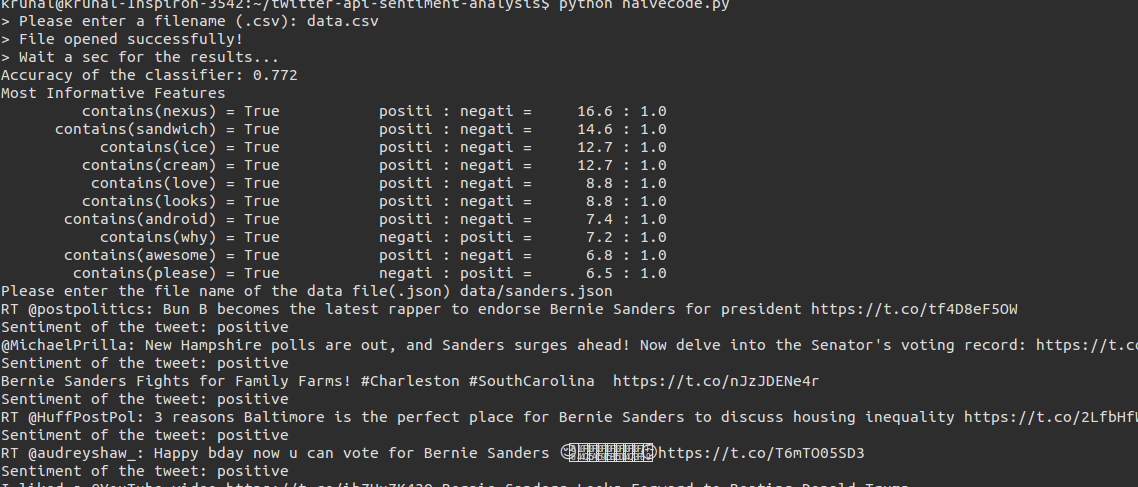
**7. TESTING**

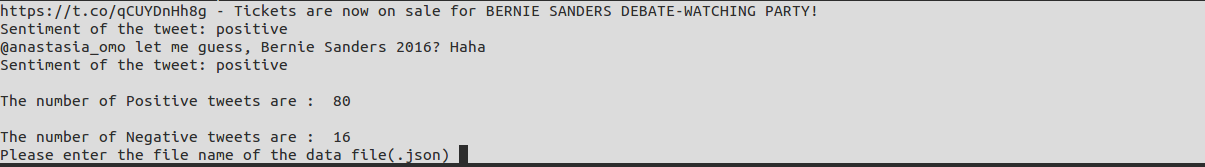
7.1 Results and Snapshots

After streaming the tweets into HDFS in real time, Naïve bayes is used in analyzing the tweets. Tweets are tagged as documents where categories are the hash tags defined in the configuration file. Later the tweets are grouped as positive, negative and neutral based on subjectivity corpus forming a dictionary of words and its polarity.









The Sentimental analysis of the tweet is shown to the user using numerical values

As shown above.

**8. CONCLUSION & SCOPE FOR**

**FUTURE WORK**

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to java twitter streaming API. It was very interesting to gather and then aggregate the social networking data so as to extract interesting patterns and recent trends from it. We got exposure to work with prominent parallel data processing tool: Hadoop. Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly.

This project helped us not only to gain knowledge about installation and configuration of hadoop distributed file system but also map reduce programming model. At the end of analysis phase data visualization was performed with the help of Google Developer. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyze sentiment, or sentiment analysis.

Twitter has lot of scope in the modern era. Twitter as a social media has many users and the number of users is increasing day by day. At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret.

**9. REFERENCES**

* <https://archive.org/download/twitter_cikm_2010>
* <http://www.cloudera.com/content/www/enus/documentation/other/tutorial/CDH5/Hadoop-Tutorial/ht_example_4_sentiment_analysis.html>
* <https://github.com/omarshammas/sentiment_analysis>
* <https://github.com/timvandermeij/sentiment-analysis/tree/master/words>
* <https://github.com/madhusudancs/sentiment-analyzer/blob/master/analyzer/train.py>
* <http://academictorrents.com/details/d8b3a315172c8d804528762f37fa67db14577cdb>
* <https://archive.org/details/twitter_cikm_2010>
* <http://www.mrgeek.me/technology/datascience/data-mining-1-5-million-tweets-for-twitter-sentiment-analysis/>
* <http://www.slideshare.net/sumit786raj/sentiment-analysis-of-twitter-data?related=1>
* <http://www.slideshare.net/niteshsinghns/twitter-sentiment-analysis-project-report>
* <http://www.alex-hanna.com/tworkshops/lesson-6-basic-sentiment-analysis/>
* <http://alexdavies.net/twitter-sentiment-analysis/>
* <http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf>