



Twitter sentiment analysis of U.S Political system

Project Report1



**By**

Lavanya Kumar Somu; ID# 14

Srikar Reddy Mallareddygari; ID# 10

Pavan Kumar Bollaram; ID# 5

Sashidhar Reddy Gowra; ID# 9

Rajasekhar Reddy Ogirala; ID# 19

October 3, 2014

Table of Contents

[Motivation 2](#_Toc400141403)

[Introduction 2](#_Toc400141404)

[Significance 2](#_Toc400141405)

[**System Architecture Design** 3](#_Toc400141406)

[**Data Model** 4](#_Toc400141407)

[Data Sources 4](#_Toc400141408)

[Algorithms 4](#_Toc400141409)

[Expected Inputs/Outputs 4](#_Toc400141410)

[**Application Specification** 5](#_Toc400141411)

[System Specification 5](#_Toc400141412)

[Activity Diagram 5](#_Toc400141413)

[Existing Applications Used: 6](#_Toc400141414)

[Design of Mobile Client 6](#_Toc400141415)

[**Documentation** 8](#_Toc400141416)

[**Project Management** 16](#_Toc400141417)

[Implementation Status Report 16](#_Toc400141418)

[**Deployment** 17](#_Toc400141419)

[Git Hub URL: 17](#_Toc400141420)

[References 17](#_Toc400141421)

# Motivation

Currently massive amount of data is rolling across various internet based applications such as Twitter, Facebook, and Amazon etc. All these applications are collecting data every minute and if we analyze this data with proper analytic techniques there will be a scope of finding social trends, favorite topics, purchasing Habits, predicting the future market areas etc.

# Introduction

Twitter is one of the most popular social media in the world. Twitter has a valuable collection of big data knowledge. Twitter is being used by all sections of the people in the society like Ordinary Citizens, Celebrities, Presidents of various countries, Different Companies, Official Government Accounts etc. Lots of valuable information is being shared through Twitter by them. By applying the right kind of data analytics on this data we can able to extract valuable insights.

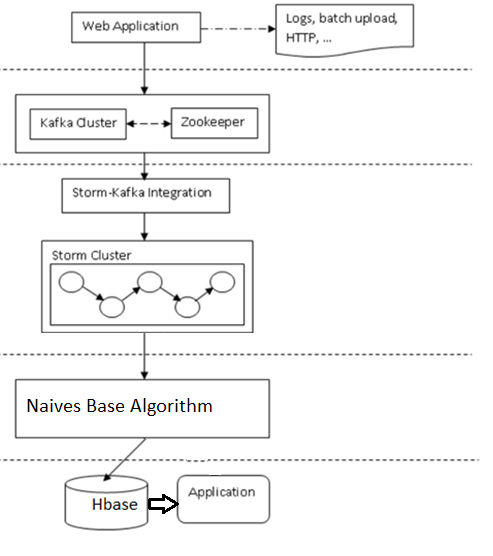
Twitters Streaming APIs are more efficient and generous unlike other API providers who poses significant API Usage limitations. So all these factors motivated my team and myself to do a Real-Time Big Data Analytics project on Twitter.

# Significance

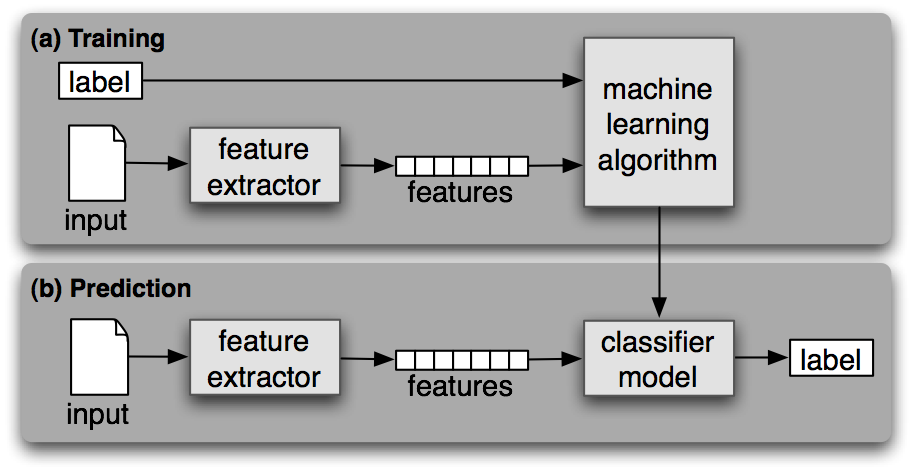
In this project we are collecting political tweets from streaming API‘s and we will be going to apply analytics on this data so that valuable and meaningful content can be retrieved and later this data can be visually represented in the form of graphs and charts for better understanding.

The primary and technical objective of this project is to get the data from twitter and predicting the sentiment towards the Republicans & Democrats.

# **System Architecture Design**



**Fig 1: Flow Chart for Twitter Tweet Analysis**

****

**Fig 2 Source:**

[**http://ravikiranj.net/drupal/201205/code/machine-learning/how-build-twitter-sentiment-analyzer**](http://ravikiranj.net/drupal/201205/code/machine-learning/how-build-twitter-sentiment-analyzer)

# **Data Model**

## Data Sources

Twitter Streaming Data: Twitter is one of the most popular social media in the world. Twitter has a valuable collection of big data knowledge.

## Algorithms

Weka Model:

The Standard approach for analyzing document classifier to transform unstructured text format datasets into a predefined or semi structured to apply a standard propositional learning schema. The Model provides a Classification Job to analyze the unstructured data. Naïve Bayes is the one of such multinomial algorithm. We have used Naïve Bayes algorithm for analyzing the JSON twitter data for mood analysis.

Sentiment analysis algorithm:

We are capturing the sentiment or mood of the tweets with the use of Naïve Bayes analysis algorithm because of the simplest mathematics involved and it is very simpler to code using programming languages. This algorithm uses the simple probabilistic classifier with naïve independence assumption. This approach is supervised learning approach and system builds their own knowledge structure with capacity of learning something new. To do classification we have to do data preprocessing, feature extraction, training and testing.

Once the data is read from the file it will be preprocessed and feature extraction will be done on the data.in this approach we will find the part of the speech of the word and we do feature extraction using pre define feature word list which intern produces feature vector. The weightage of the tweet sentiment is evaluated based on the feature vector.

## Expected Inputs/Outputs

* **Input:** Twitter real time Streaming Dataset
* **Output:**  GUI of Twitter Analytics(Android Graphs & Charts)

# **Application Specification**

## System Specification

* **Cloudera Manager:**

Cloudera Manager is one of the best application management environment for Apache Hadoop. Cloudera manager provides the Hadoop eco system which is useful to Manage, Monitor, diagnose and integrate the analytics applied on Big Data sets.

* **Storm:**

Storm is a real time distributed computation system. It is useful for real time processing of streaming data.

* **Weka:**

The Standard approach for analyzing document classifier to transform unstructured text format datasets into a predefined or semi structured to apply a standard propositional learning schema. The Model provides a Classification Job to analyze the unstructured data. Naïve Bayes is the one of such multinomial algorithm. We have used Naïve Bayes algorithm for analyzing the JSON twitter data for mood analysis.

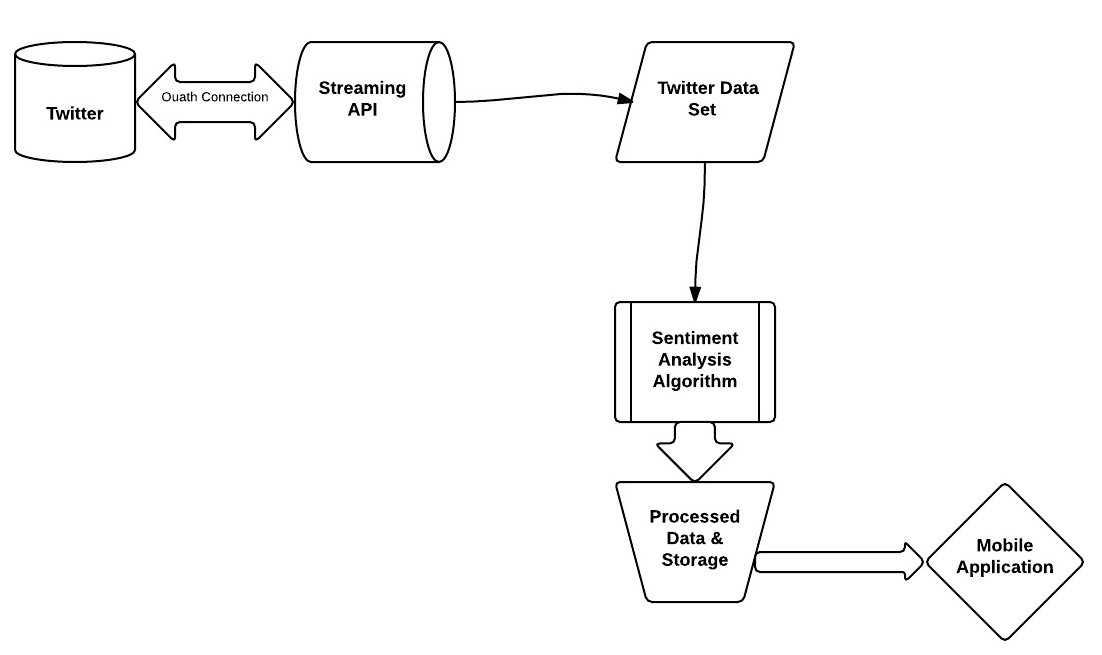
* **Hbase:**

HBase is useful to apply random reads and writes on real-time data. HBase is a non-relational DB model. It provides capability to create billions of rows and millions of columns for distributed data storage.

* **Eclipse ADT:**

ADT is a plugin for Eclipse IDE. Eclipse ADT is useful to build Android applications.

## Activity Diagram



## Existing Applications Used:

1. **Name:** habakkuk-alpha

**Description:** Habakkuk is an application for streaming tweets.

**URL:** *https://github.com/telvis07/habakkuk-alpha.git*

1. **Name:** Weka Naives Base Algorithm

**Description:** Feature Extraction & Sentient Analysis

**URL:**

*http://preciselyconcise.com/apis\_and\_installations/training\_a\_weka\_classifier\_in\_java.php*

## Design of Mobile Client

Mobile Application:

Developed a mobile application which reads the collected twitter dataset. User can check the content of the data set and at the same time after sentimental analysis user can check the each tweet and their respective analysis whether it reflects positive mood or negative mood.at the end we are displaying the report to user on click of button which included on mobile app.

Features:

1. Reads the data from the HBase by establishing the connection.
2. Analyses mood of the data (tweets) retrieved from HBase.
3. Displays the data in pie-chart graphical view with the different colors.

Technologies:

Used Java and Android development kit for coding and GUI of mobile client

A chart Engine for graphical view

GUI:

GUI is contains Text view, web view and buttons on screen. Used Pie chart for visualization of the analysis.

Implementation

* Implemented logic to establish connection to Hbase and trying to fetch the reat time data which is stored in HBase.
* Once the data retrieved we are finding the sentiment of the mood by analyzing the data collected.
* The required data which is needed for graphical representation is passed to AchartEngine client.
* Created a category series and render objects and passed these objects to chart engine to create pie chart.

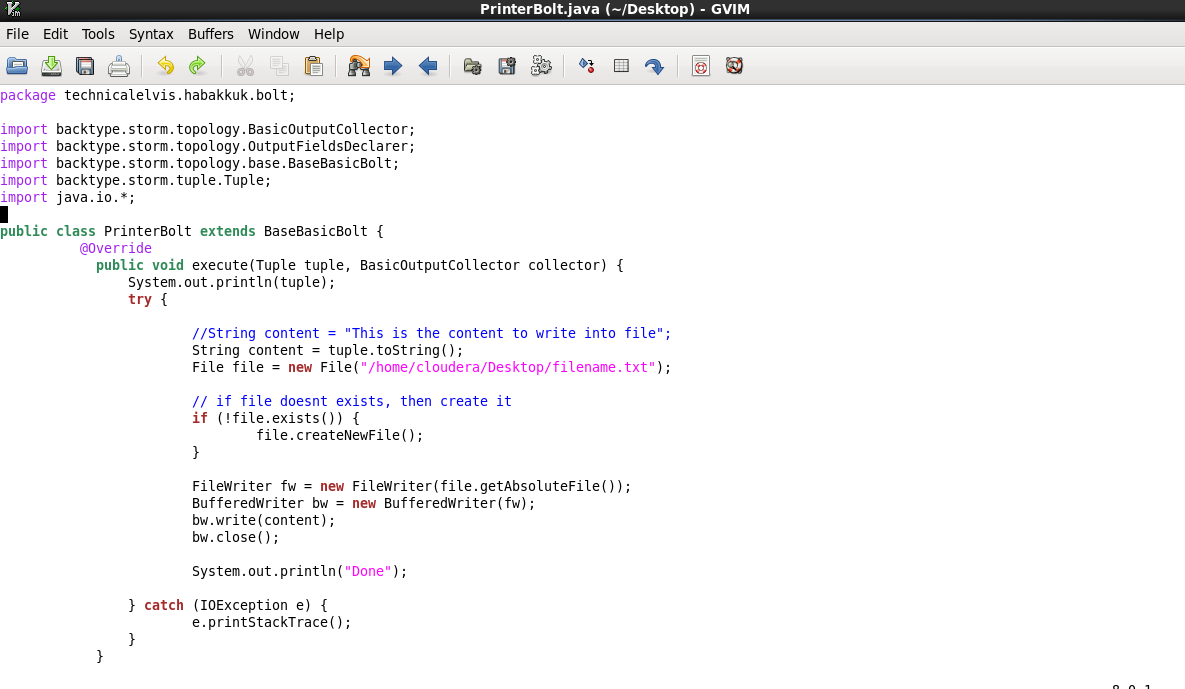
Designed layout which includes buttons and text view and displaying the rendered output on screen using pie chart.

# **Documentation**

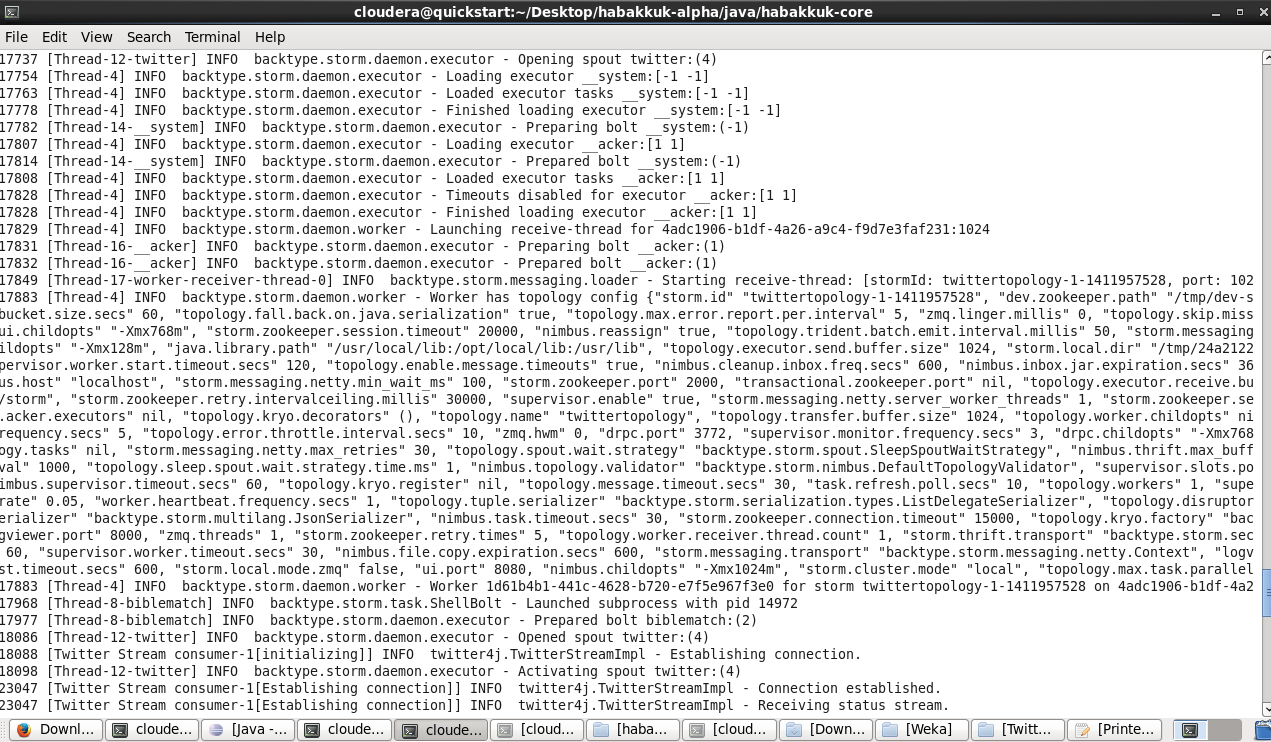
* Get a Habakkuk project from github from below command.

git clone <https://github.com/telvis07/habakkuk-alpha.git>

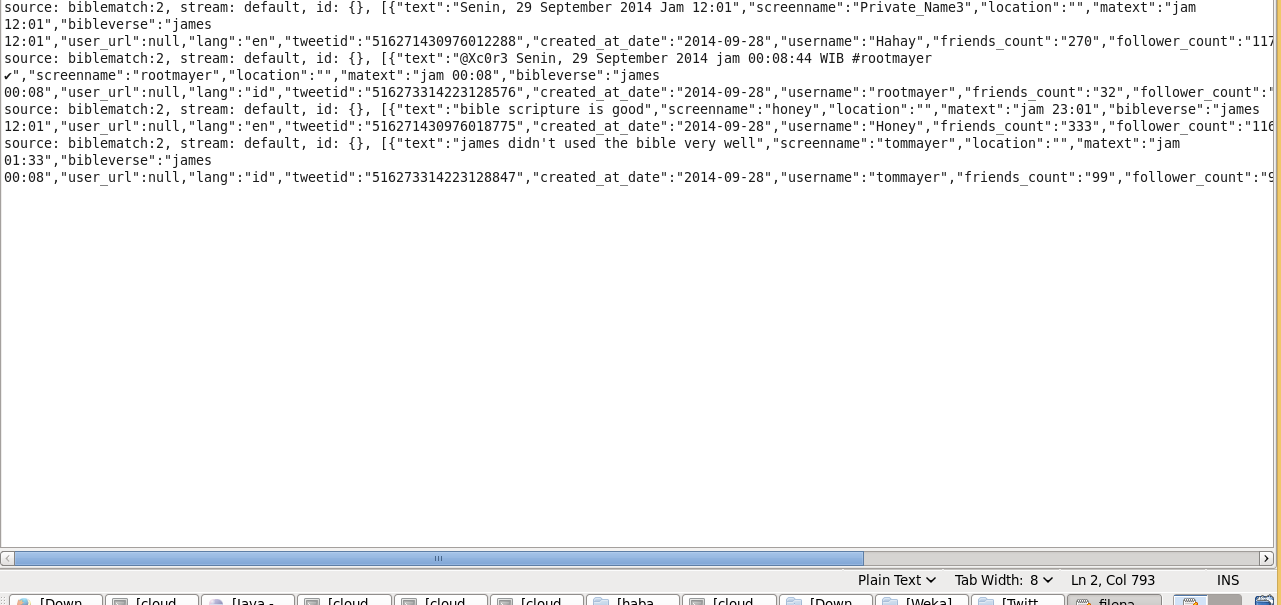
* Modify the multi lang resources file located in habakkuk-alpha/java/habakkuk-core/multilang/resources
* Modify the Habakkuk.properties file for accessing the twitter streaming API by adding the API Keys & tokens of twitter application
* Modify the printerBolt.java code to capture the twitter streaming in local machine by adding below code to the existing logic.



* Run the storm for streaming the twitter data as shown in below screenshot.

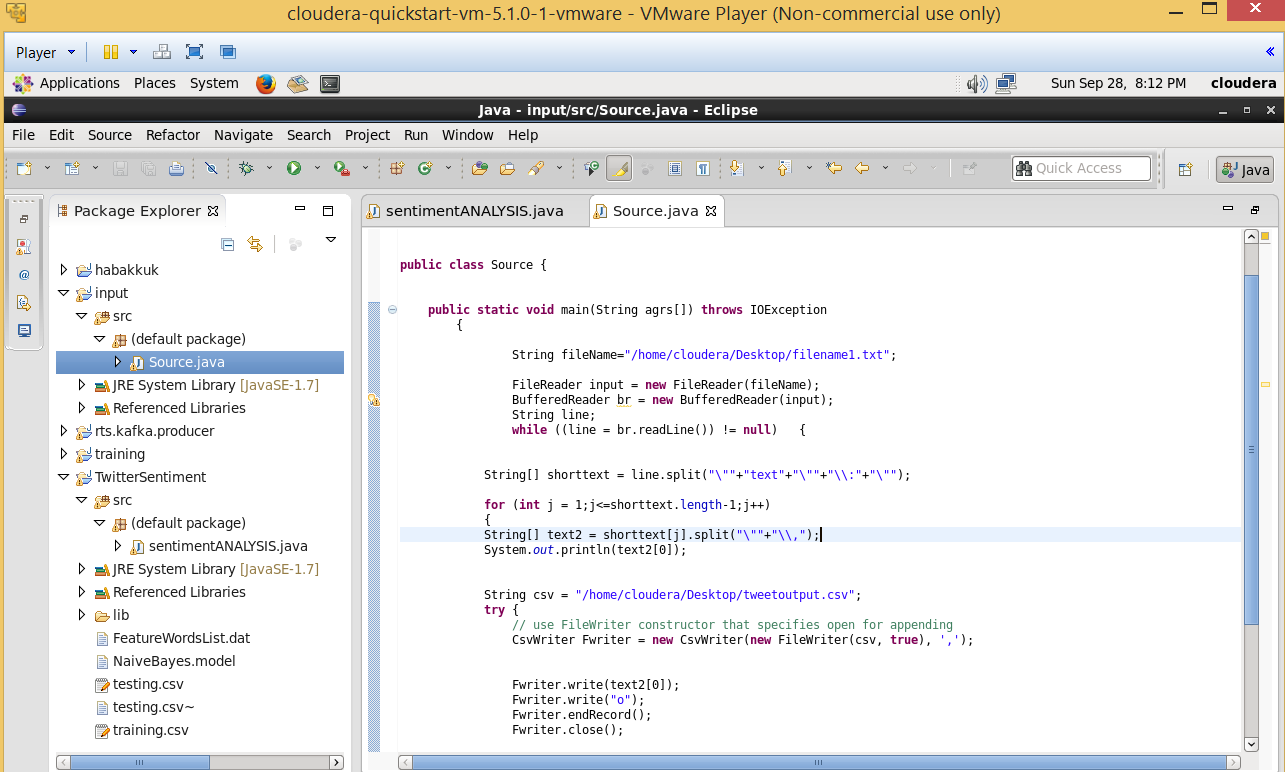


**Twitter Streaming File:**



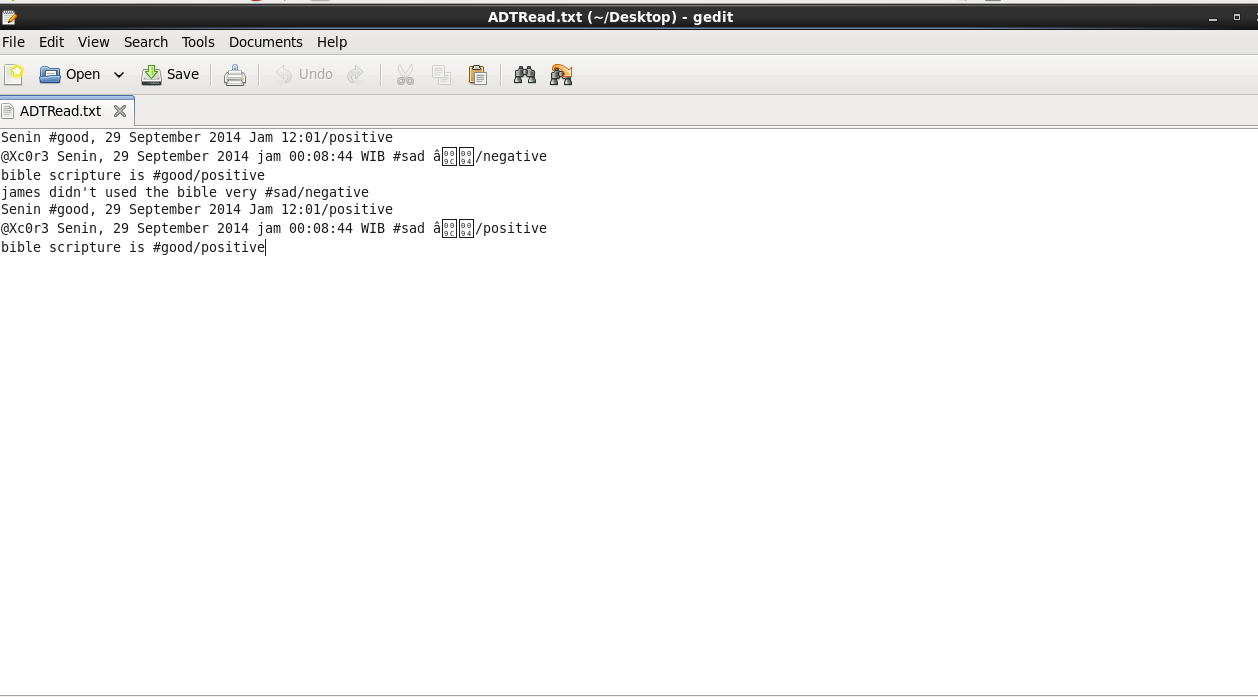
**Extracting tweets from JSON Data:**

A Source.java class is written to extract the tweets from above streamed data using below screenshot.



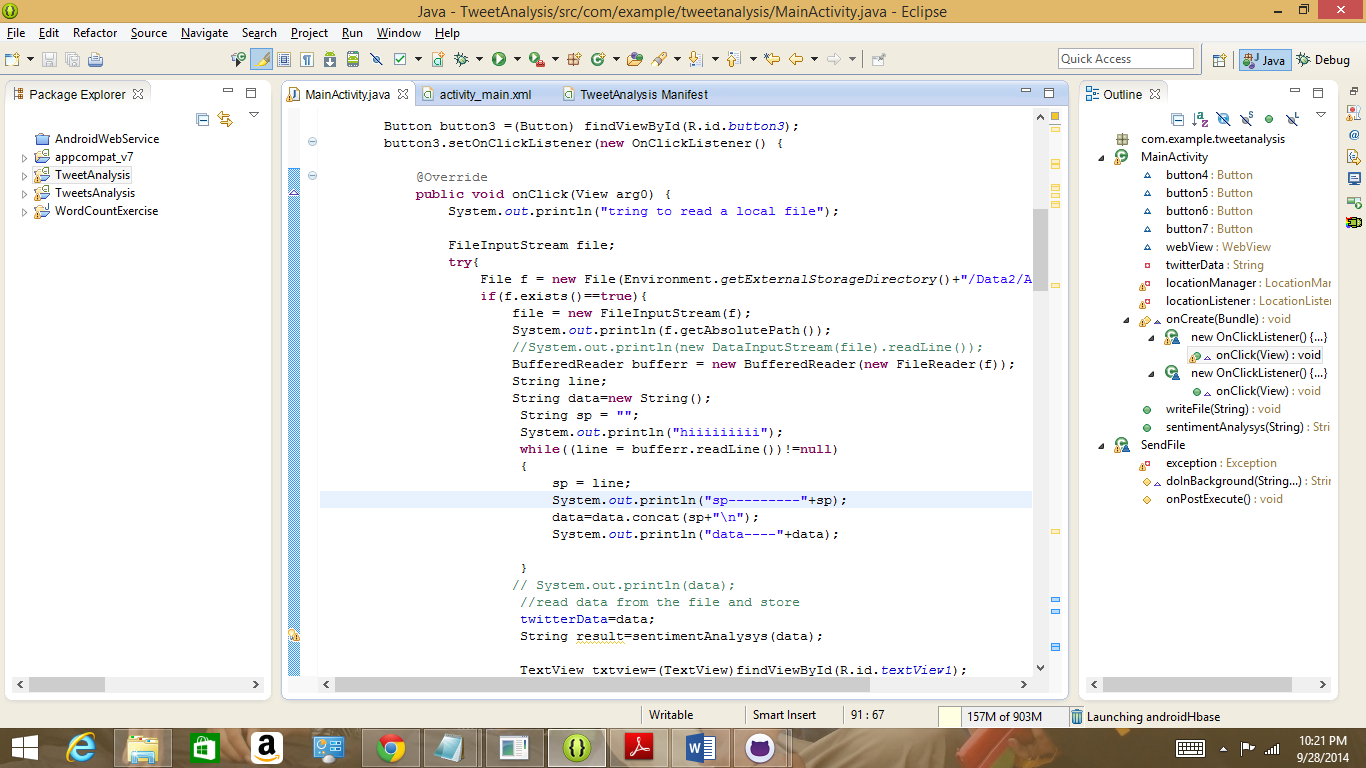
**Processed File Using Weka & Navies Algorithm:**

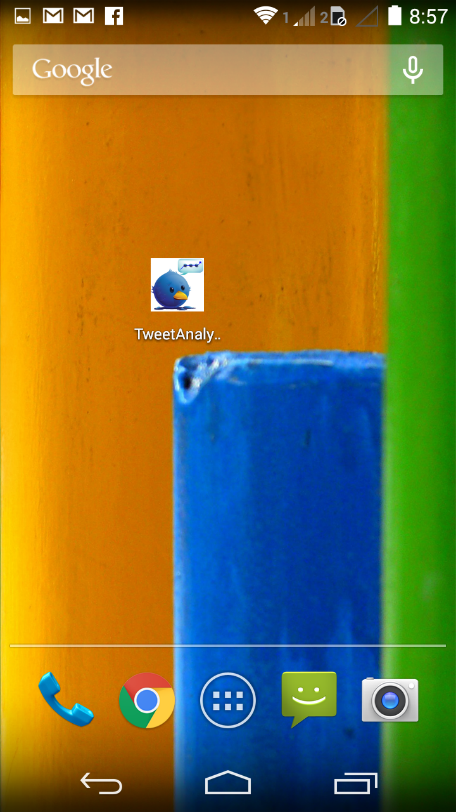
Sentiment of above tweets is calculated using the Navies algorithm. The output is stored in local machine as shown in below screenshot.



**Mobile Application Design & Output:**

A Mobile Application is developed to analyze and view the data from the above step. Please find the below screen shots for code & application.

  
Android App Icon:



Android App:



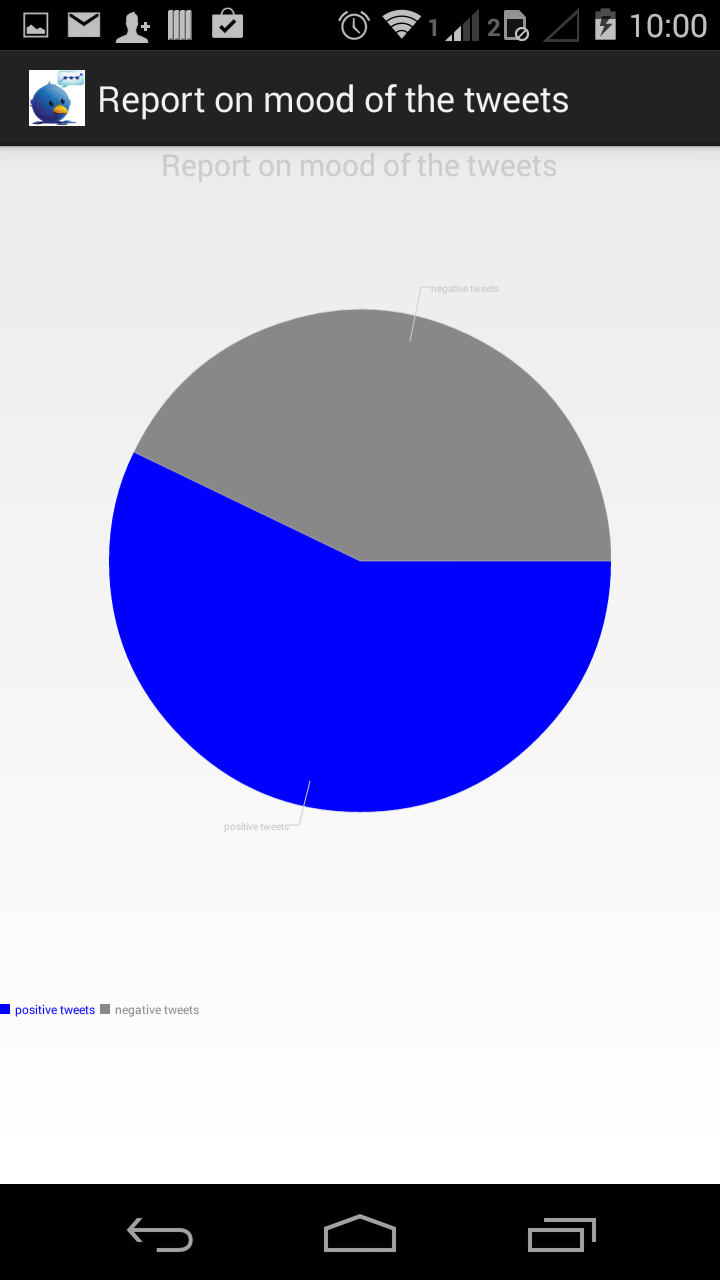
View Sentiment Analysis Button:



Generate Report Button:



View Chart Button:



# **Project Management**

## Implementation Status Report

* **Work Completed:**
* Extracting Streaming Twitter Data
* Storing a Data into a local machine using java class.
* Analyzing the tweet sentiment using the navies algorithm
* Designing a mobile application to read the output file to display charts.

**Responsibility:**

Srikar Reddy Mallareddygari – Streaming Twitter Data, Hbase Design, Documentation – 20%

Pavan Kumar Bollaram – Android ADT – 20%

Somu Lavanya Kumar – Applying Sentiment Analysis, Hbase Design, Documentation – 20%

Sashidhar Reddy Gowra – Extracting Tweet Text data from Tweets, Analytics – 20%

Rajasekhar Reddy Ogirala - Extracting Tweet Text data from Tweets – 20%

* **Work to be Completed:**
* Extracting a real time streaming data based on a location and keywords.
* Developing a Web Service to be used by Android GUI.
* Automate the process of Android GUI to retrieve the data from a centralized HBase Location.
* Improving the Accuracy for Sentiment of twitter tweets.
* **Issues/Concerns:**
* Storm Streaming connectivity.
* Hbase Server Issue.

# **Deployment**

## Git Hub URL:

<https://github.com/pavankumar-b/Project-CS590RA>

# References:

<http://www.sysomos.com/docs/Inside-Twitter-BySysomos.pdf>

Symeon Papadopoulos, Yiannis Kompatsiaris, "Social Multimedia Crawling for Mining and Search," Computer, vol. 47, no. 5, pp. 84-87, May, 2014.

<http://ieeexplore.ieee.org.proxy.library.umkc.edu/stamp/stamp.jsp?tp=&arnumber=6818959&tag=1>

<http://www.alex-hanna.com/tworkshops/lesson-6-basic-sentiment-analysis/>

<http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-hadoop/>

<http://cran.r-project.org/web/packages/jsonlite/vignettes/json-apis.html>

<http://www.slideshare.net/Hadoop_Summit/realtime-sentiment-analysis-app-using-hadoop-and-h-base>

<https://github.com/twitter/twitter4j>

<http://hortonworks.com/wp-content/uploads/2014/05/Hortonworks.BusinessValueofHadoop.v1.0.pdf>

[*https://dev.twitter.com/streaming/overview*](https://dev.twitter.com/streaming/overview)

[*https://github.com/telvis07/habakkuk-alpha.git*](https://github.com/telvis07/habakkuk-alpha.git)

*http://preciselyconcise.com/apis\_and\_installations/training\_a\_weka\_classifier\_in\_java.php*

[*https://apps.twitter.com/*](https://apps.twitter.com/)

*blog.csdn.net/hljlzc2007/article/details/13275441*

*http://www.csvreader.com/java\_csv\_samples.php*