**CAPSTONE PROJECT REPORT**

(Project Term January 2018-December 2018)

## A NOVEL APPROACH ON SENTIMENT ANALYSIS FOR TWITTER

Submitted by

**Name: Registration Number:**

**Perumalla Murali Mallikarjuna Rao 11501280**

**Baddigam Mahesh Reddy 11503509**

**Swetha R 11500954**

**Swathy R 11500956**

**Apurva Roshan 11505703**

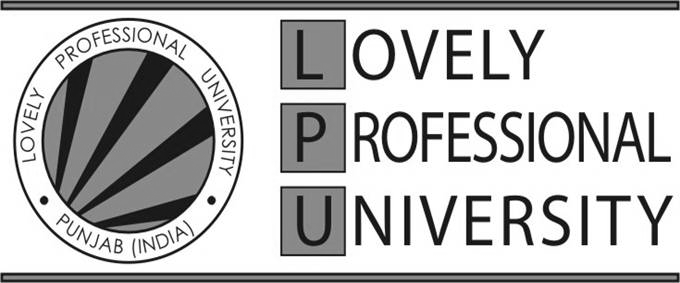
**Project Group Number: CSERGC0171**

**Course Code: CSE439**

Under the Guidance of

**(Mr. Amritpal Singh-Assistant Professor)**

# School of Computer Science and Engineering



**TOPIC APPROVAL PERFORMA**

School of Computer Science and Engineering (SCSE)

**Program:** P192-ND: Integrated B.Tech. - M.Tech. (Computer Science & Engineering)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **COURSE CODE:** | CSE439 | **REGULAR/BACKLOG:** | Regular | **GROUP NUMBER:** CSERGC0171 |
| **Supervisor Name**: | Amritpal Singh | **UID:** 17673 |  | **Designation:** Assistant Professor |

**Qualification:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **Research Experience:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SR.NO.** | **NAME OF STUDENT** | **REGISTRATION NO** | **BATCH** | **SECTION** | **CONTACT NUMBER** |
| 1 | Apurva Roshan | 11505703 | 2015 | K1519 | 9411884370 |
| 2 | Baddigam Maheshreddy | 11503509 | 2015 | K1507 | 8500979365 |
| 3 | Perumalla Murali Mallikarjuna Rao | 11501280 | 2015 | K1518 | 7382891098 |
| 4 | Swathy R | 11500956 | 2015 | K1518 | 8113030592 |
| 5 | Swetha R | 11500954 | 2015 | K1518 | 8113030592 |

**SPECIALIZATION AREA**: Database Systems **Supervisor Signature:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**PROPOSED TOPIC**: A Novel Approach for sentiment analysis for Twitter

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Qualitative Assessment of Proposed Topic by PAC** | | | |
| **Sr.No.** | **Parameter** | | | **Rating (out of 10)** |
| 1 | Project Novelty: Potential of the project to create new knowledge | | | 6.40 |
| 2 | Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students. | | | 6.40 |
| 3 | Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program. | | | 6.80 |
| 4 | Project Supervision: Project supervisor’s is technically competent to guide students, resolve any issues, and impart necessary skills. | | | 7.20 |
| 5 | Social Applicability: Project work intends to solve a practical problem. | | | 6.80 |
| 6 | Future Scope: Project has potential to become basis of future research work, publication or patent. | | | 6.60 |
| **PAC Co** | | **mmittee Members** | | |
| PAC Member 1 Name: Kewal Krishan | | UID: 11179 | Recommended (Y/N): Yes | |
| PAC Member 2 Name: Raj Karan Singh | | UID: 14307 | Recommended (Y/N): Yes | |
| PAC Member 3 Name: Sawal Tandon | | UID: 14770 | Recommended (Y/N): Yes | |
| PAC Member 4 Name: Robin Prakash Mathur | | UID: 14597 | Recommended (Y/N): Yes | |
| DAA Nominee Name : Kuldeep Kumar Kushwaha | | UID: 17118 | Recommended (Y/N): Yes | |

**Final Topic Approved by PAC: A Novel Approach for sentiment analysis for Twitter**

**Overall Remarks:** Approved

**PAC CHAIRPERSON Name:** 11024: Amandeep Nagpal **Approval Date:** 09 May 2018

**DECLARATION**

We hereby declare that the project work entitled “A NOVEL APPROACH TO TWITTER SENTIMENT ANALYSIS ” is an authentic record of our own work carried out as requirements of Capstone Project for the award of B. Tech degree in Computer Science Engineering from Lovely Professional University, Phagwara, under the guidance of Mr. Amritpal Singh, during August to November 2018. All the information furnished in this capstone project report is based on our own intensive work and is genuine.

Project Group Number:CSERGC0171

Name of Student 1: Perumalla Murali Mallikarjuna Rao

Registration Number:11501280

Name of Student 2: Baddigam Mahesh Reddy

Registration Number:11503509

Name of Student 3: Swetha R

Registration Number: 11500954

Name of Student 4: Swathy R

Registration Number: 11500956

Name of Student 4: Apurva Roshan

Registration Number: 11505703

Date:

Date:

Date:

Date:

Date:

**CERTIFICATE**

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Capstone Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Capstone Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in Computer Science Engineering from Lovely Professional University, Phagwara.

**Signature and Name of the Mentor**

**Designation:**

**School of Computer Science and Engineering,**

Lovely Professional University,

Phagwara, Punjab.

Date:

**ACKNOWLEDGEMENT**

Completing a task is never a one-man effort. It is the results of valuable contribution of a number of individuals in a direct or indirect manner that helps you shape and achieve an objective.

This project would not have taken shape, without the guidance provided by Mr. AMRITPAL SINGH, our Mentor, who helped us in our project and resolved all the technical problems and also helped in understanding technical aspects of the project. We profusely thank them for the support provided to us.

We also express a deep sense of gratitude for providing us the opportunity and trusting us for their project. We are very grateful.

**Table of Contents:**

1. Introduction

2. Profile of the Problem. Rationale/Scope of the study (Problem Statement)

3. Existing System

* Introduction
* Existing Software
* DFD for present system
* What’s new in the system to be developed

4. Problem Analysis

* Product definition
* Feasibility Analysis
* Project Plan

5. Software Requirement Analysis

* Introduction
* General Description
* Specific Requirements

6. Design

* System Design
* Design Notations
* Detailed Design
* Flowcharts

7. Testing

* Functional testing
* Structural testing
* Levels of testing
* Testing the project

8. Implementation

* Implementation of the project
* Conversion Plan
* Post-Implementation and Software Maintenance

9. Project Legacy

* Current Status of the project
* Remaining Areas of concern
* Technical and Managerial lessons learnt

10. User Manual: A complete document (Help Guide) of the software developed***.***

11. Source Code (where ever applicable) or System Snapshots

12. Bibliography

1. **INTRODUCTION**

With fast development in user of Social Media lately, the specialist gets pulled in towards the utilization of web based life information for suppositions examination of individuals or specific item or individual or occasion. Twitter is one of the broadly utilized online networking stage to express the considerations.

This project is based on Twitter Sentiment analysis using machine learning algorithms in python. In this project we used Tweepy API for authentication and streaming tweets form the twitter. Collected tweets are Pre-processed with the help of Natural Language Processing and regular expressions. With the help of Machine learning algorithms, we have created a model for sentiment analysis in which we make use concept of pickling in python. With the help of this model we can predict the sentiment of the tweets. In this project we have developed an application with the help of PySide2, purpose of the application is to clean and analyse the sentiment of the tweets taken from user.

We will be comparing different tools like Spark and Python for the better accuracy of analysis. Visualization is the better way to memorize and to understand everything easily. Visualization helps a layman to understand everything at a glance. In this project we used some popular online and offline visualization tools for the visualization of our results. Those tools are IBM Watson/Congo analytics, Tableau, python, R studio.

To make the project more interactive and user friendly we developed a GUI using python which has the ability to extract the data and to visualize the extracted data manually.

* 1. **Tweepy API**

Tweepy is an Application Programming interface provided as a tweepy library in python. Tweepy library provides OAuthHandler module for authentication to twitter. With the help of Twitter client and Streaming module in tweepy we have done streaming on the topics we need. In order to connect with twitter we need the twitter credentials.

* + 1. **Tweet Structure in json format**

**Tweet in json format:**

**“**{"created\_at":"Mon Nov 05 18:55:47 +0000 2018","id":1059519663523184640,"id\_str":"1059519663523184640","text":"@austinmcg5 my pixel died dude I am so sorry. I'm getting the 3!! I didn't mean to ignore you :(","display\_text\_range":[12,96],"source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","truncated":false,"in\_reply\_to\_status\_id":1059519363370360834,"in\_reply\_to\_status\_id\_str":"1059519363370360834","in\_reply\_to\_user\_id":551036534,"in\_reply\_to\_user\_id\_str":"551036534","in\_reply\_to\_screen\_name":"austinmcg5","user":{"id":488890084,"id\_str":"488890084","name":"melancholy mane","screen\_name":"AdamFromThePack","location":null,"url":null,"description":"Oh yeah\nHere come the painbirds","translator\_type":"none","protected":false,"verified":false,"followers\_count":333,"friends\_count":253,"listed\_count":6,"favourites\_count":9710,"statuses\_count":13875,"created\_at":"Fri Feb 10 23:32:02 +0000 2012","utc\_offset":null,"time\_zone":null,"geo\_enabled":true,"lang":"en","contributors\_enabled":false,"is\_translator":false,"profile\_background\_color":"ACDED6","profile\_background\_image\_url":"http:\/\/abs.twimg.com\/images\/themes\/theme18\/bg.gif","profile\_background\_image\_url\_https":"https:\/\/abs.twimg.com\/images\/themes\/theme18\/bg.gif","profile\_background\_tile":false,"profile\_link\_color":"038543","profile\_sidebar\_border\_color":"FFFFFF","profile\_sidebar\_fill\_color":"F6F6F6","profile\_text\_color":"333333","profile\_use\_background\_image":true,"profile\_image\_url":"http**”**

The key attributes are the following:

* **text**: the text of the tweet itself
* **created\_at**: the date of creation
* **favorite\_count, retweet\_count**: the number of favourites and retweets
* **favorited, retweeted**: boolean stating whether the authenticated user (you) have favourited or retweeted this tweet
* **lang**: acronym for the language (e.g. “en” for english)
* **id**: the tweet identifier
* **place, coordinates, geo**: geo-location information if available
* **user**: the author’s full profile
* **entities**: list of entities like URLs, @-mentions, hashtags and symbols
* **in\_reply\_to\_user\_id**: user identifier if the tweet is a reply to a specific user
* **in\_reply\_to\_status\_id**: status identifier id the tweet is a reply to a specific status
  1. **Twitter Application**

In order to Authenticate with twitter, we need create a developer account in twitter which is Twitter Apps. This Application generates access keys and secret keys used to authenticate and stream the tweets.

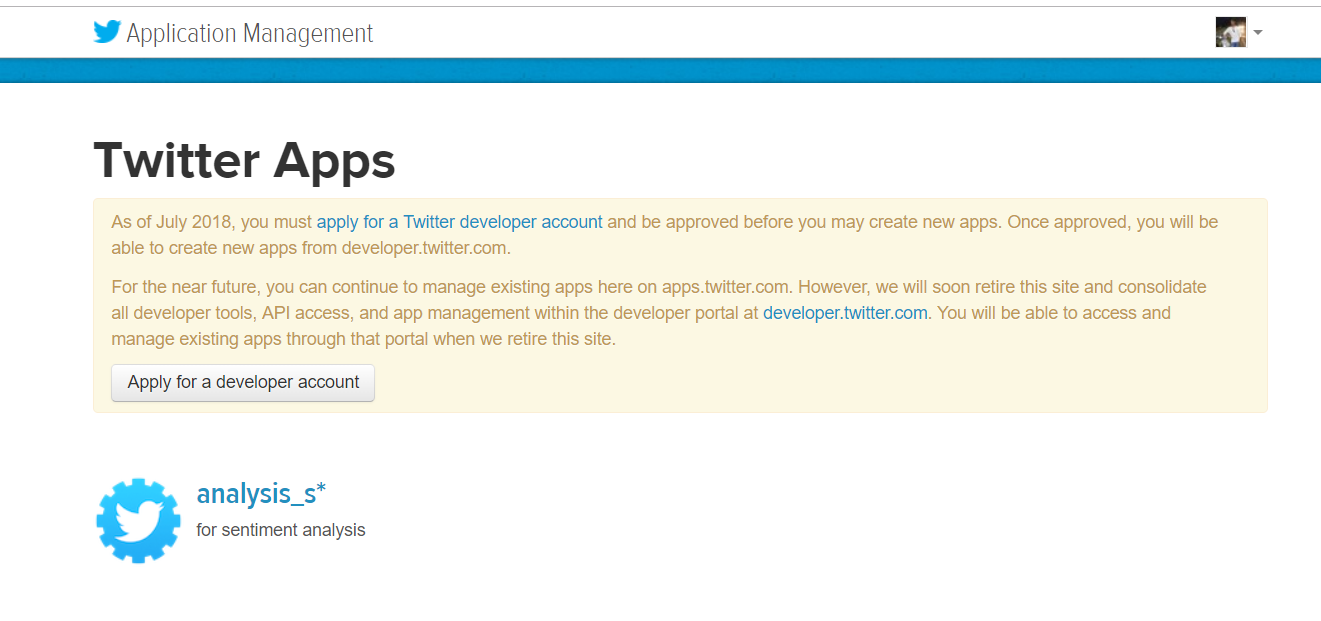


fig1.0 This is our twitter application that we have used

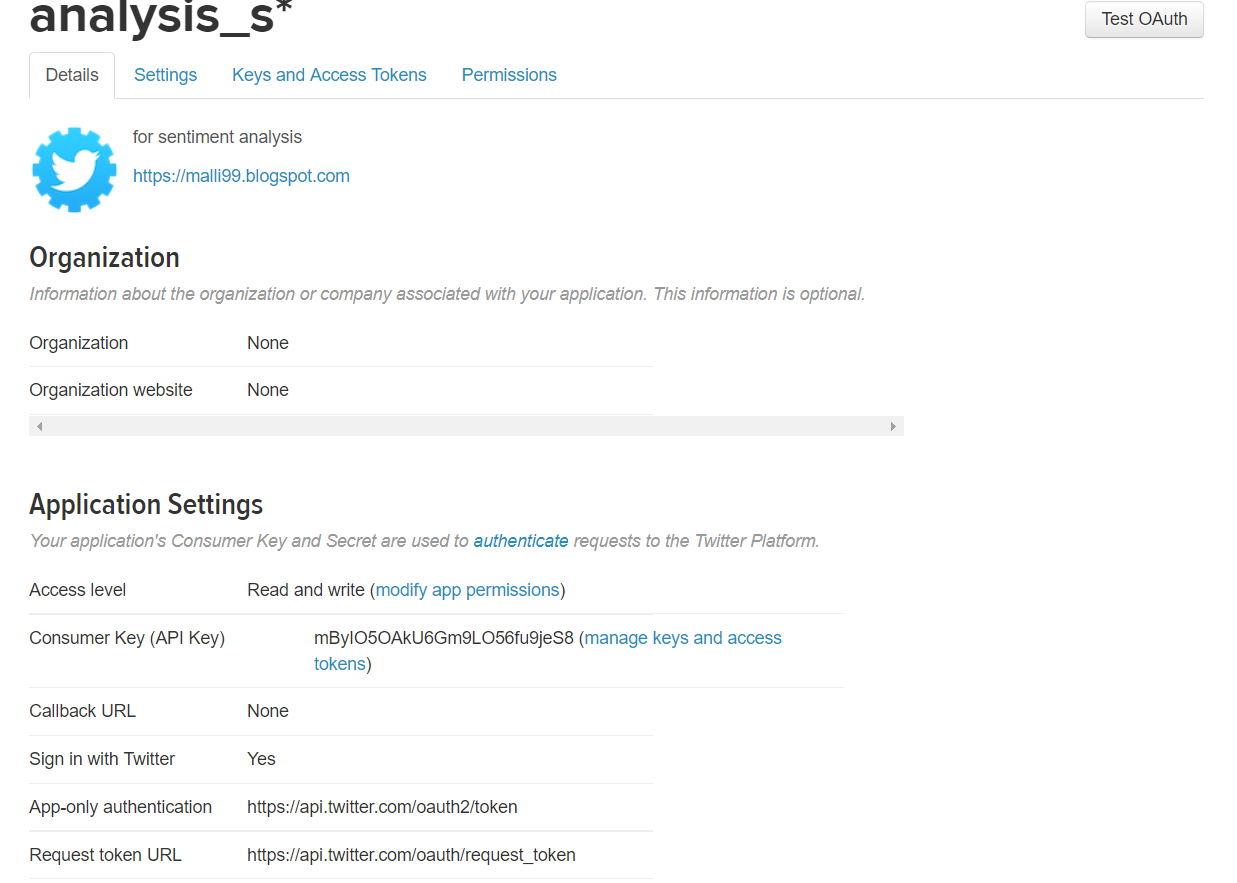


fig1.0.1 This is screenshot of twitter application

* 1. **Sklearn Library**

Sklearn library provides machine algorithms in python. In this project we have used few machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector machines. We also made use of Count Vectorizer and tfidf vectorizer for converting the text into features which is helpful in training the machine learning model.

* + 1. **Naive Bayes Algorithm**

This algorithm is one best machine learning algorithm used for test classification. We have used Multinomial NB as classifier. This algorithm works on the probability. With the help of this algorithm we have classified the tweets into three classes i.e. positive, negative and neutral. Naive Bayes algorithm is supervised learning algorithm based on applying Bayes theorem.

**P(y | x1,……,xn) = P(y)P(x1,…. xn | y)**

**P(x1,….., xn)**

## Logistic regression

We used logistic regression to build the model. It predicts the probability of occurrence of an event by fitting data to a logit function. Logistic Regression is a classification algorithm used to predict a binary outcome (1/0, Yes/true, True/False) given a set of independent variables.

Equation of Logistic Regression:

**Log(p/1-p)= β0  + β(Age)**

* + 1. **SVM**

SVM stands for Support Vector Machines. SVMs are set of supervised learning methods used for classification. SVC, NuSVC , LinearSVC are classes capable of performing multiclass classification on a dataset. Different Kernal Functions can be specified for the decision function.

table1.3.0 shows the accuracy of the algorithm used on hate speech detection dataset taken from internet

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Multinomial Naive Bayes algorithm | 81 % accuracy |
| Logistic regression algorithm | 92 % accuracy |
| linearSVC algorithm | 96% accuracy |
| NuSVC algorithm | 77%accuracy |
| SGDClassifier algorithm | 71% accuracy |
| SVC algorithm | 55% accuracy |

Table 1.3.0

* + 1. **Dataset Creation**

We have gathered tweets on 3 products iPhone R, one plus 6t, pixel 3xl. we have created a data set of the tweets. We have gathered total 12 thousand tweets which took nearly 3days. With the help of the dataset we have analysed the sentiment of these people on these products and popularity of these products.

### **Hate Speech detection**

We have taken an external dataset for hate speech detection in tweets. Using our system, we have cleaned the dta and analysed the data and classified the data into different sentiment class. with the help of the world clouds we have visualized the top positive and negative tweets which says about hate speech in tweets

* + 1. **Hash Tag Analysis:**

In this project we have collected the hashtags present in tweets gathered with the help of regular expressions and made some visualizations on it.

* 1. **GUI Application**

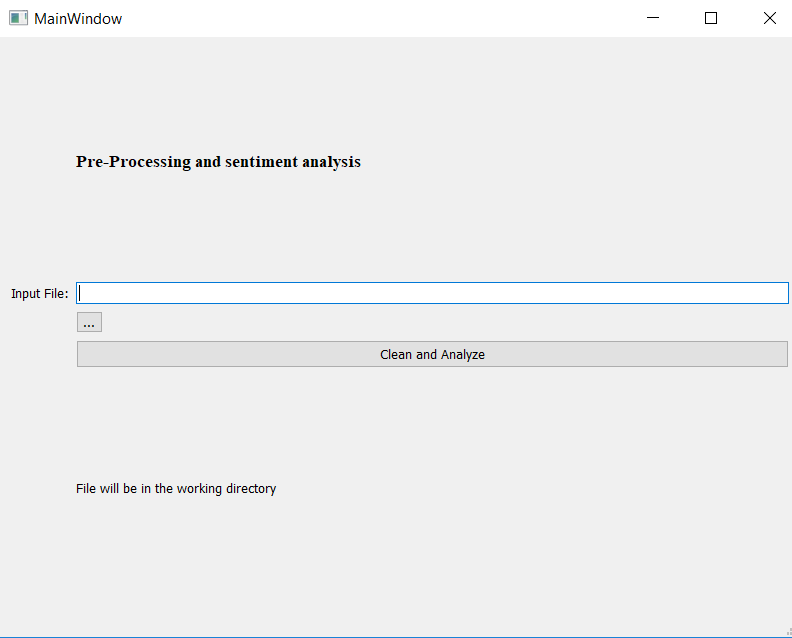
 We have created GUI application using pyside2 package in python. our GUI application is for offline cleaning and sentiment analysis on a dataset provided by user.

fig 1.4.0 Figure shows the GUI interface.

* 1. **Real Time sentiment analysis on python console**

We can analyse the sentiment in run time that is real time analysis in python console. With the help of the pickling concept we have saved a naïve Bayes model used to classify the sentiment at the runtime .and also, we can detect the hashtags if any present in the tweet at runtime. we can also print the cleaned tweet at run time but it is bit time consuming.

# SCOPE OF STUDY

The study covers the need and effectiveness of Novel approach for sentimental analysis. Our model or approach saves precious time in cleaning the tweets and analysing the sentiment. Sentiment analysis can be helpful in hate speech detection, know about the views of people on specific products.

# EXISTING SYSTEM

## Introduction

The existing system is a system for tweets streaming and pre-processing the gather tweets and analysing the sentiment. System uses the feature extraction methods for converting the text into features used for building the machine learning model used for sentimental analysis. system uses the pickling concept for saving the ML model. We have gathered tweets on different products and created dataset on these products.

## Specifications of system

* + 1. **Twitter Streaming:**

Using the API, we can authenticate to twitter. we can also stream the tweets based on twitter screen name or any keyword or hashtag mention by the user. Streamed data will be of json format which is similar to the dictionary.

**3.2.2 Pre-Processing:**

The data gather in Jason format is not easily understandable and text of tweet contains many special characters and hashtags, stop words, links, emoticons which has to be removed. In pre-processing we clean the tweets and store them in csv file which can be further used for analysis.

* + 1. **Sentiment Analysis:**

We have used textblob for sentiment analyse. then we used this data to train the machine learning model. with the help the trained machine learning model we can classify the test data. ML model is used to predict the sentiment of the tweet and perform classification.

**3.2.3** **Visualization**

In visualization part, we visualized all the results in the form of graphs using various platforms like IBM Watson, Python, R Studio. For the further visualization, we created a GUI using Python.

* + 1. **Comparison**

Further we compared the accuracy of various platforms. Mainly focused on the accuracy of Python and Spark. Came into a clear cut idea that …….. is far better than the ………… because it yields the accuracy by ………

## Data Flow Diagrams

Data flow diagram (DFD) represents the flow of data between different processes in a system. It is a graphical technique that depicts information flow and the transformations that are applied as the data moves from input to output. It provides a simple, intuitive method for describing business processes without focusing on the details of computer systems. Data Flow Diagram is an attractive technique that provides what users do rather than what computers do.

DFDs only involve four symbols. They are as following:

* Process

 Transform of incoming data flow(s) to outgoing flow(s).

* Data Flow

 Movement of data in the system

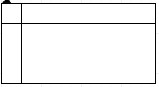
* Data Store

Data repositories for data that is not moving. . It may be as simple as buffer or a queue or as sophisticated as a relational database.

* External Entity

Sources of destinations outside the specified system boundary.

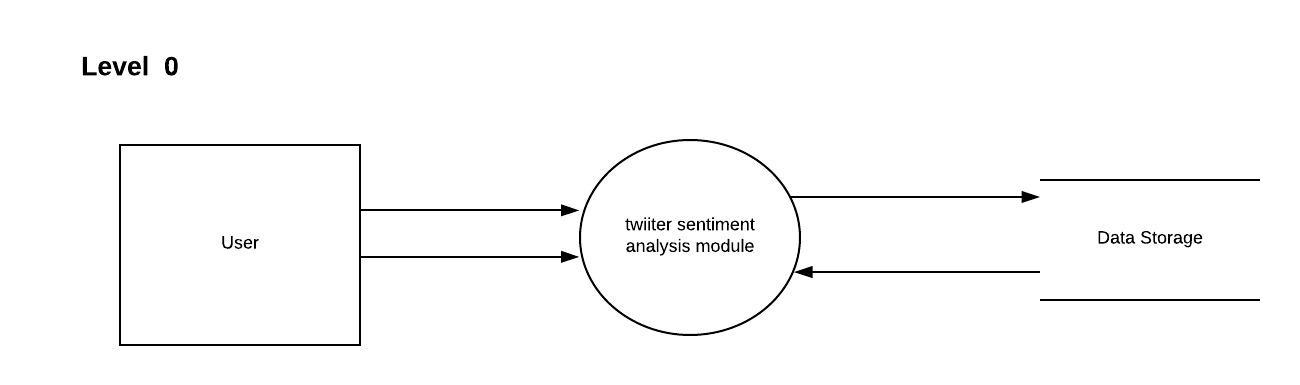
* Internal Storage



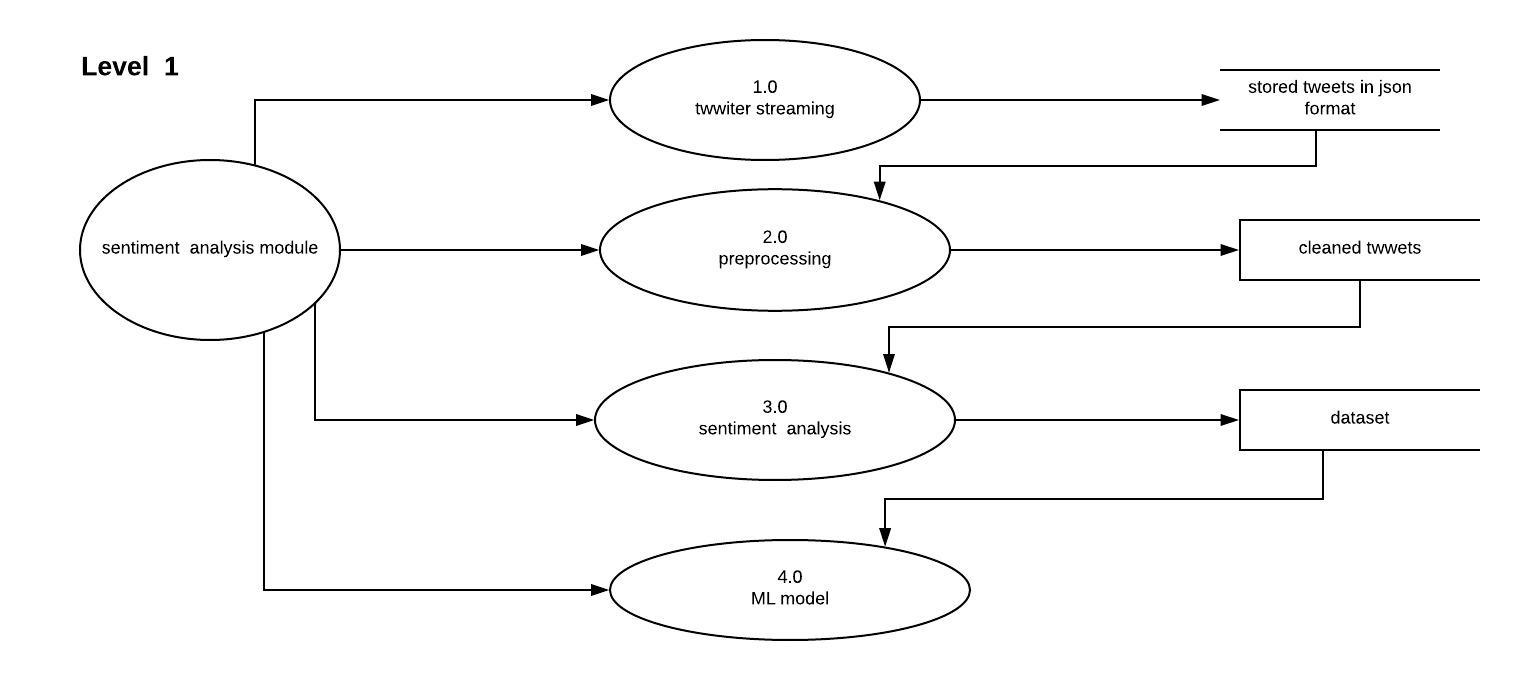
* **Relationships and Rules**: The Data Flow Diagram may be used for any level of data abstraction. DFD can be partitioned into levels, Each level has more information flow and functional details than the previous level. Highest level is Context Diagram. Some important points are:

1. Data arrows show input and output.
2. Data Stores NOT shown. They are within the system.
3. 1 bubble (process) represents the entire system.

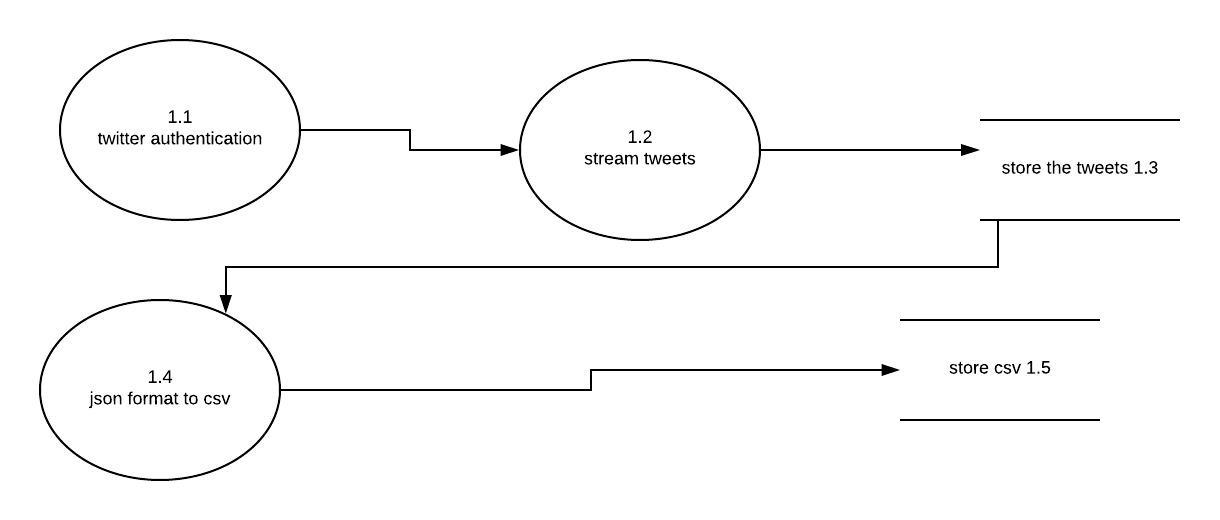
* **DFD at Level 0**



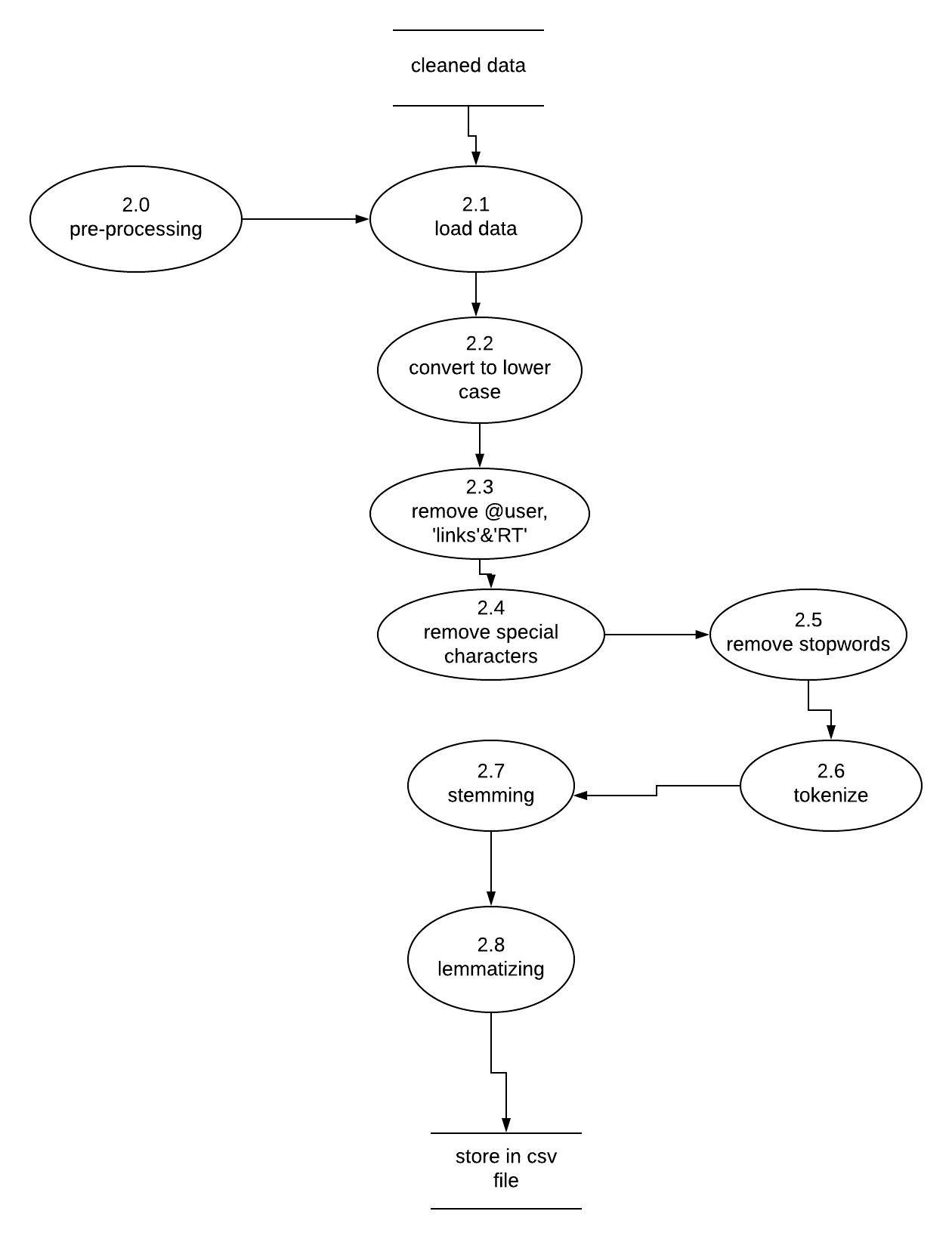
* **DFD at Level 1**



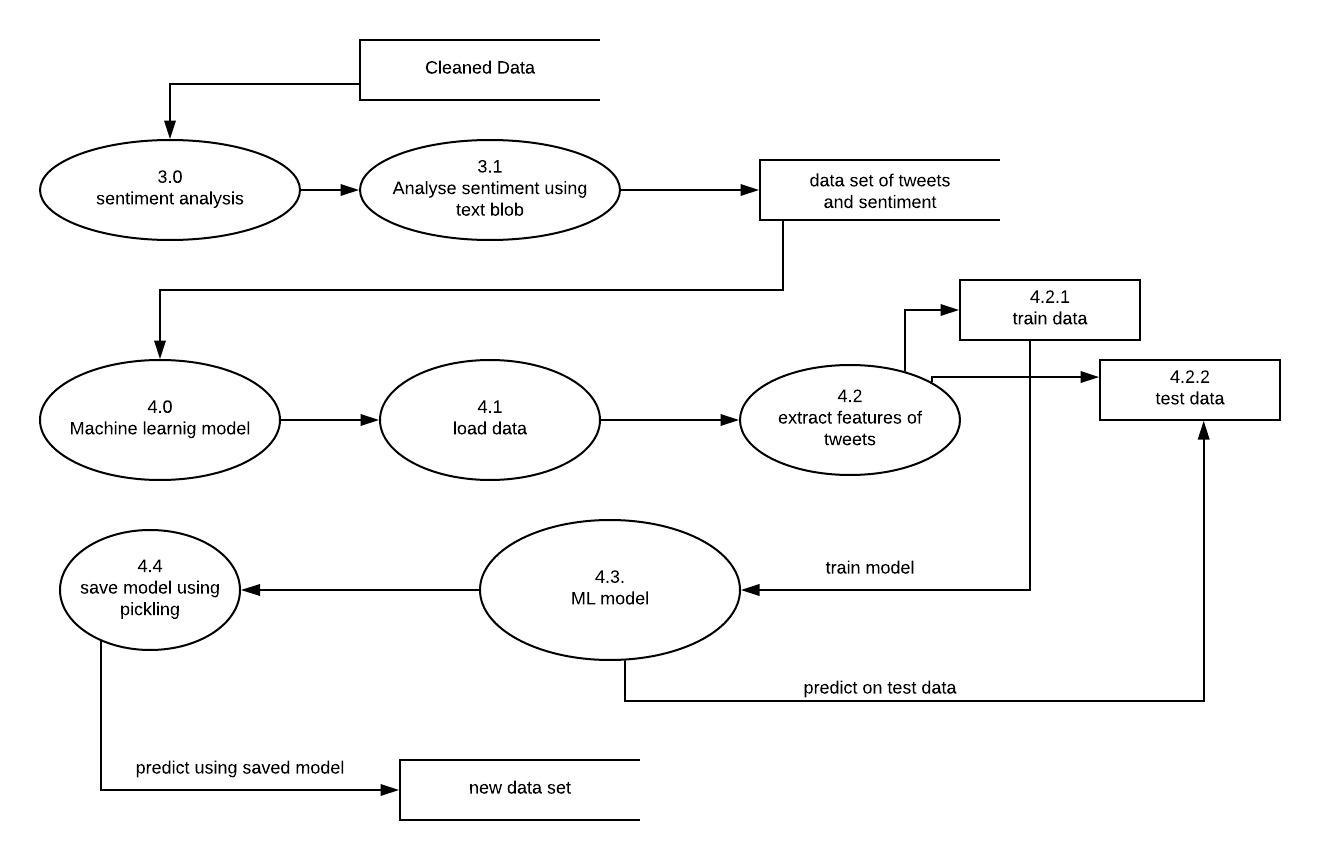
* **DFD at Level 2**



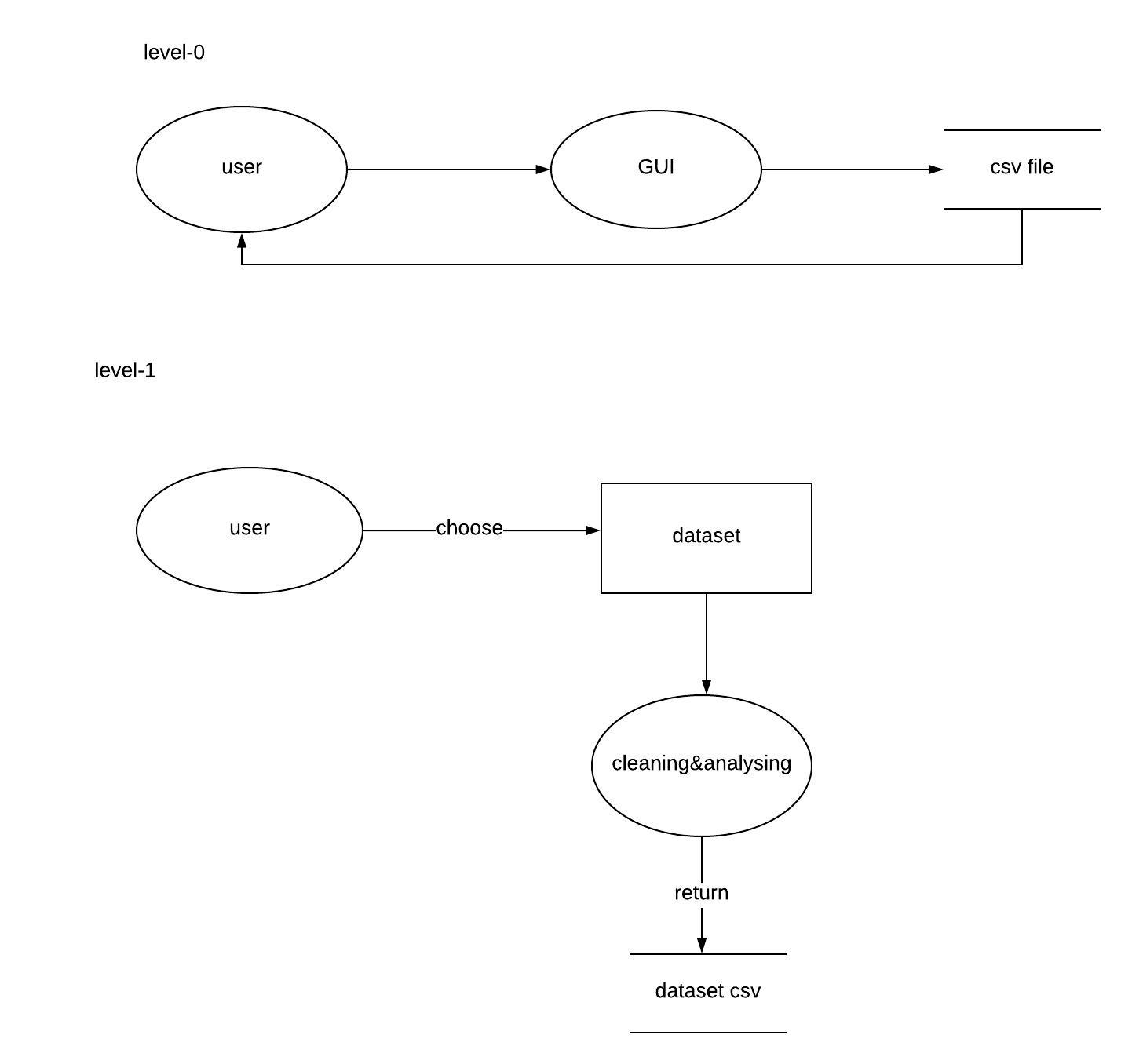
* **DFD at level 2 for pre-processing**

****

* **DFD at level 2 for ML model**



* **DFD for GUI**

****

# 4. PROBLEM ANALYSIS

In this phase, the main problem concern about collection of tweets, cleaning the data and building a machine learning model to analyze the positive, negative and neutral tweets. At last, we developed methods which can make user friendly.

**4.1 Product definition**

Well, our project is all about to perform twitter streaming, pre-processing of tweets, extraction tweets and perform sentiment analysis on any product or any famous personalities. We did sentimental analysis on mobiles which are entered into the market like One plus 6T, iPhone XR, Google pixel 3xl. This project will show how people are thinking about the mobiles which we mentioned above i.e., positive, negative or neutral. We used machine learning algorithms, natural language processing tools.

Our project includes following features:

* Ability to Authenticate with twitter with ease
* Can perform preprocessing on any data provided by user
* Machine learning model for sentiment analysis
* Gathering the insights for the tweet data
* Visualizing of the data

**4.2 Feasibility Study**

The feasibility study proposes one or more conceptual solutions to the problem set for project. The conceptual solution gives an idea of what the new system will look like. Feasibility study defines what will be done on system and what will remain manual. It was found that proposed system is technically, operationally and economically feasible. Also, it is necessary to consult the users to see if the computerized data recording function satisfies the user objectives and can be fitted into current system operation. The major outputs of this phase are technical feasibility report, cost- benefit analysis report, operational feasibility report, and time schedule of the project.

**4.2.1 Technical Feasibility:**

* Project is user friendly
* Less RAM is required.
* Hardcoding of content is minimized to use less memory of the system

**4.2.2 Economic Feasibility:** It is a cost benefit analysis keeping in view the system is economically feasible. It is so due to following points:

* No need to make a survey on each mobile individually
* No need to waste huge amount of money by giving to third party agencies
* It’s very cost effective customers can come to a conclusion which mobile was doing good in market
* Mobile companies can come to decision what kind of features people really want.
* By seeing all the above things mobiles companies may enhance features and sales of their mobiles.

**4.2.3 Operational Feasibility:** Operational feasibility is the measure of how well a proposed project solves the problems and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of project development.

* As the system is user friendly, the system is well linked and approved by the user’s demands and more facilities can also be provided.
* The system has been prepared keeping in mind the user’s current requirements as well as change in future. The proposed system will not cause any harm to the user but instead, it will enhance the result in a better way.

## 4.3 Project Plan

A project plan is produced as management activities commence. The plan defines the process and tasks to be conducted, the people who will do the work, and the mechanisms for assessing risks, controlling change, and evaluating quality. To develop this project, we followed Evolution Model in which our requirements could change during development only. We set our schedule according to hours. As we spent few hours on project daily, we got less than 120 days to complete it. Our schedule is as follow:

1. Problem Identification- 10 days.
2. Feasibility Study- 5 days.
3. Requirement Analysis and Installation- 10 days.
4. Design

* Physical Design- 10 days
* Database Design- 5 days

1. Coding Implementation

* Database Implementation - 23 days
* Front- End Implementation- 15 days.
* Connectivity between front-end and back-end- 15 days.

1. Testing- 5 days.
2. Finalizing report- 5 days.

**4.3.1 Assignment of Work**

|  |  |
| --- | --- |
| **TEAM MEMBER NAME** | **WORK ASSIGNED** |
| * **Murali mallikarjuna rao perumalla** | * Twitter App and Authentication and tweets streaming * Json to csv conversion * Machine learning algorithms and pickling concept * Streaming and anlysis of gathered tweets on 3 mobile products * Real time sentiment analysis on python console * documentation |
| * **Mahesh Reddy Baddigam** | * Natural language processing * Cleaning using NLP and regular expressions * Hate speech detection on data set taken from internet * Added one more machine learning algorithm * Documentation |
| * **Swathy R** | * Visualisations using IBM Watson Analytics, IBM Congo Analytics * Visualisations using R * Comparison of spark with python using naïve bayes classifiers algorithm * IBM Visualization GUI * Documentation * Research Paper for the following for future Scope |
| * **Swetha R** | * Sentimental Analysis Using R * Visualization Using R * Comparison of Visualization tools used * Comparison of spark with python using logistic regression algorithm * Documentation * Research Paper for the following for future Scope |
| * **Apurva Roshan** | * Visualization Using Matplotlib in Python Using GUI * Visualization Using Tableau * Developed the Connecting GUI interface of Processing and Visualization * Comparison of spark with python using k-means clustering and random forest algorithm * Documentation * Research Paper for the following for future Scope |

# 5. SOFTWARE REQUIREMENT ANALYSIS

In this project, we used a functional Software Requirement, functional means providing particular service to the user.

## 5.1 Introduction to tools and platforms

This project package is developed using tools for development which is best suited for development of this project package. These tools are as follows:

* **Python:** Python is an object-oriented programming language similar to PERL, which has gained popularity due to its syntax and readability. Python is relatively easy to learn and transport, which means that its instructions can be interpreted in different operating systems, including UNIX, Mac OS, MS-DOS, OS / 2 and several versions of Microsoft Windows 98. It was created by Guido van Rossum, a former resident of the Netherlands, whose favorite comedy group at that time was the Monty Python Flying Circus. The source code is freely available and open for modification and reuse. Python has a large number of users.
* **Tweepy:** An easy to use python library to access the twitter API.
* **TextBlob:** TextBlob is a Python library (2 and 3) for the processing of textual data. It provides a simple API to immerse yourself in common natural language processing tasks, such as partial voice dialing, nominal phrase extraction, sentiment analysis, classification, translation and more.
* **Sklearn:** It includes several classifications, regression, and grouping algorithms, including support vector machines, random forests, gradient augmentation, k-averages, and DBSCAN, and is designed to interact with Python NumPy digital and scientific libraries. and SciPy.
* **Pandas:** Pandas is a Python package that provides fast, flexible and expressive data structures, designed to make work with "relational" or "tagged" data easy and intuitive. It claims to be the fundamental element of the high level to perform a practical and real-world data analysis in Python.
* **Numpy**: NumPy is a Python package used for scientific computing. The NumPy package is used to perform several operations. The ndarray (NumPy Array) is a multidimensional matrix used to store values of the same type of data. These matrices are indexed as sequences, starting with zero.
* **Matplot**: Matplotlib is a crawl library for Python. It is used with NumPy to provide an environment that is an effective alternative to open source code for MatLab. It can also be used with graphic tools such as PyQt and wxPython.
* **PySide2**: It is used for making GUI application in python.
* **Re**: A regular expression in a programming language is a special string of text that is used to describe a search pattern. It is extremely useful to extract text information such as code, files, records, spreadsheets or even documents.
* When using the regular expression, the first thing we must recognize is that everything is essentially a character and we write patterns to match a specific sequence of characters, also called a string. The ASCII or Latin letters are those of their keyboards and Unicode is used to match the foreign text. Includes numbers, punctuation and all special characters like $ # @ !!%, etc.
* **NLTK:** The Natural Language Toolbox (NLTK) is a platform that is used to create Python programs that use data in human language for an application in the statistical processing of natural language. It contains text processing libraries for the creation of tokens, analysis, classification, acquisition, labeling and semantic reasoning.
* **Wordcloud:** Word Cloud is a data visualization technique that is used to represent text data in which the size of each word indicates its frequency or importance. Important textual data points can be highlighted using a word cloud. Word clouds are widely used to analyze data from social networking websites.

**5.2 Specific Requirements**

**5.2.1 Functional Requirements**

The functional requirement is twitter application.

**5.2.2 Requirements for twitter application**

The requirements for the twitter application.

**General**

**Login:**

**Functional requirement:**

* **Description:** User is able to login to apps.twitter.com using e-mail and password.
* **Processing:** e-mail and password will be authenticated.
* **Output:** Users are logged in to twitter to make an application.

**Register:**

**Functional requirement:**

* **Description:** User is able to register to application using e-mail and password.
* **Processing:** e-mail and password are authenticated.
* **Output:** Users are logged in to twitter.

**5.2.3 Requirements for streaming and pre-processing**

**Functional Requirement: Authentication**

* **Description:** User need consumer key, consumer secret key, access token, access token secret key.
* **Processing:** tweepy is used for authentication with twitter.
* **Output:** twitter authentication.

**Functional Requirement: Streaming**

* **Description:** Used for extracting tweets from twitter.
* **Processing:** tweepy is used for streaming.
* **Output:** tweets in json format.

**Functional Requirement: Storing tweets**

* **Description:** Storing the tweets in local storage.
* **Processing:** file handling is used to store tweets into file.
* **Output:** tweets stored in text file

**Functional Requirement: Gathering tweets from json**

* **Description:** extracting text of tweet from json format.
* **Processing:** Pandas is used to extract the tweets from json format.
* **Output:** tweets stored into csv file.

**Functional Requirement: Cleaning**

* **Description:** used NLTK and RE tools.
* **Processing:** done with tokenizing, stemming, lemmatizing
* **Output:** cleaned tweets

**Functional Requirement: Sentimental analysis**

* **Description:** analyzed sentimental of the tweets based on the polarity.
* **Processing:** textblob library has been used
* **Output:** sentiment of tweets

**5.2.4 Requirements for Machine learning model**

**Functional Requirement: feature extraction**

* **Description:** converting the tweets in to features for ML model.
* **Processing:** countvectorizer() and tfidf-vectorizer has been used from sklearn library
* **Output:** features

**Functional Requirement: ML model**

* **Description:** build and training the model using train dataset
* **Processing:** ML algorithms from sklearn
* **Output:** trained model

**Functional Requirement: testing**

* **Description:** testing the model using test dataset
* **Processing:** ML algorithms from sklearn
* **Output:** tested model

**Functional Requirement: prediction**

* **Description:** predicting the sentiment of the tweets
* **Processing:** using ML model
* **Output:** prediction values

**5.2.5 Requirements for visualization**

**Functional Requirement: visualization**

* **Description:** dataset has been to visualize
* **Processing:** matplotlib is used to visualize the data
* **Output:** plots

**6. DESIGN**

**6.1 System Design:**

Twitter Streaming

* Authentication
* Streaming
* Storing tweet data

Pre-Processing

* Load uncleaned tweet Data
* Clean Data
* Store cleaned tweet Data

Sentiment Analysis using text blob

* Load tweet data
* Analyse sentiment.polarity of tweet
* Store in csv file

Machine Learning model

* Get data set
* Convert text into features
* Split into train and test set
* Train the model using train set
* Test the model using test set
* Save the trained model using pickle
* Use saved trained model for prediction of new data

Visualization

* Load data
* Using matplotlib or platforms or frameworks like tableau, IBM Watson for visualising the data
* Plot chart
* Word Cloud
* Comparison of Visualization tools

## Design Notations

A diagram of the sequence of movements or actions of people or things involved in a complex system or activity or a graphical representation of a computer program in relation to its sequence of functions (as distinct from the data it processes).

• **Start/End Symbol:** The terminator symbol marks the starting or ending point of the system. It usually contains the word "Start" or "End."



• **Action or Process Symbol:** A box can represent a single step ("add two cups of flour"), or an entire sub-process ("make bread") within a larger process.



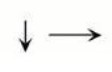
• **Decision Symbol:** A decision or branching point. Lines representing different decisions emerge from different points of the diamond.



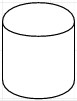
* **Input/output Symbol:** Represents material or information entering or leaving the system, such as customer order (input) or a firms (output).



• **Arrows:** It indicates the direction and order of program execution.



• **Cylinder:** It indicates the volatile database used in the application.



• **Curved Rectangle:** It represents the already stored data in the applications which cannot be modified by user.



• **Double Faced Rectangle:** It represents the predefined process.



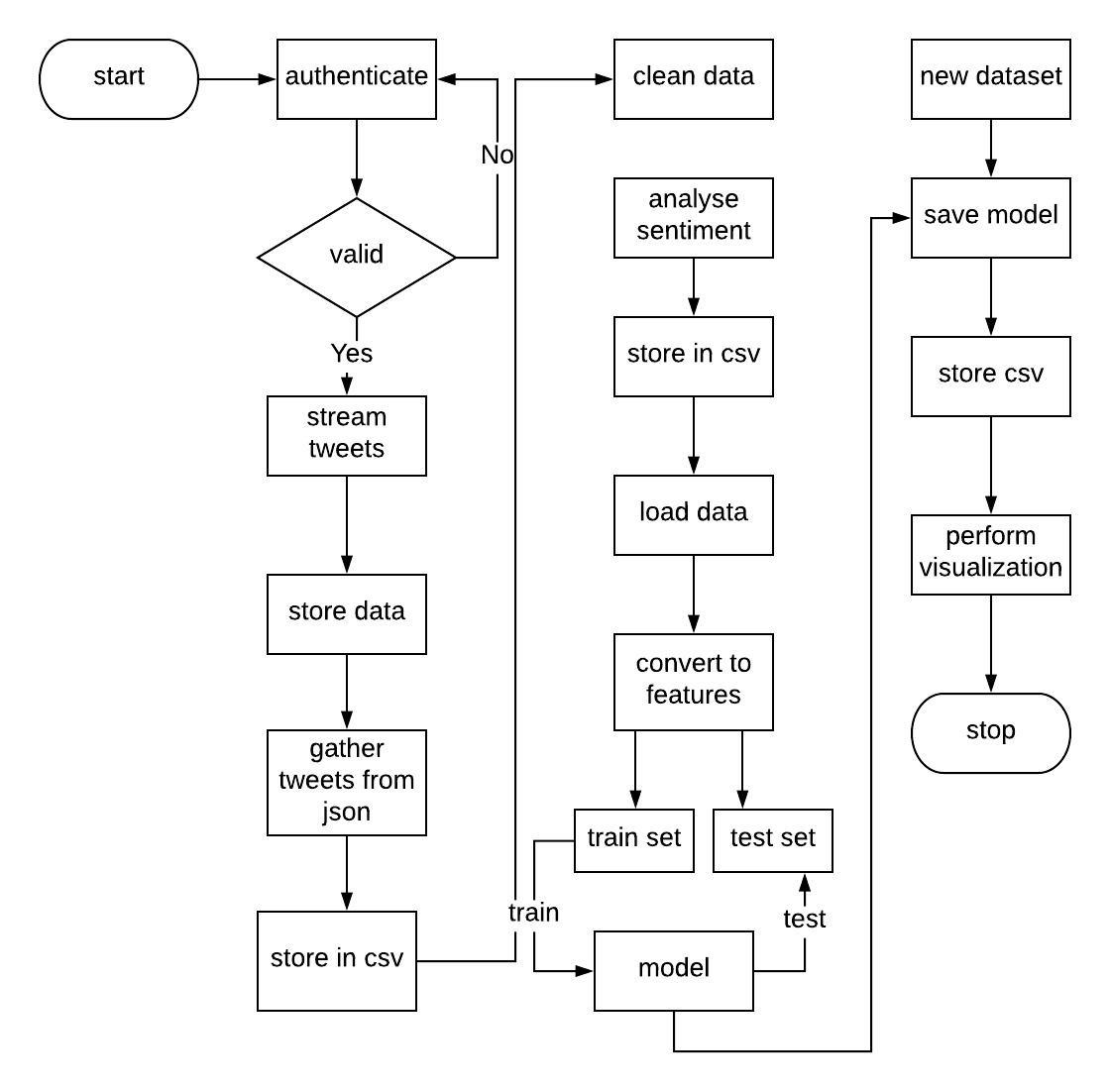
## Detail Design

* It’s simple and understanding interface will make it easy for anyone for best doctor assistance application.
* This application can easily be operating anywhere any time.
* Interact with Doctor/Patient/ Attendant.
* Capable of downloading any medical reports.

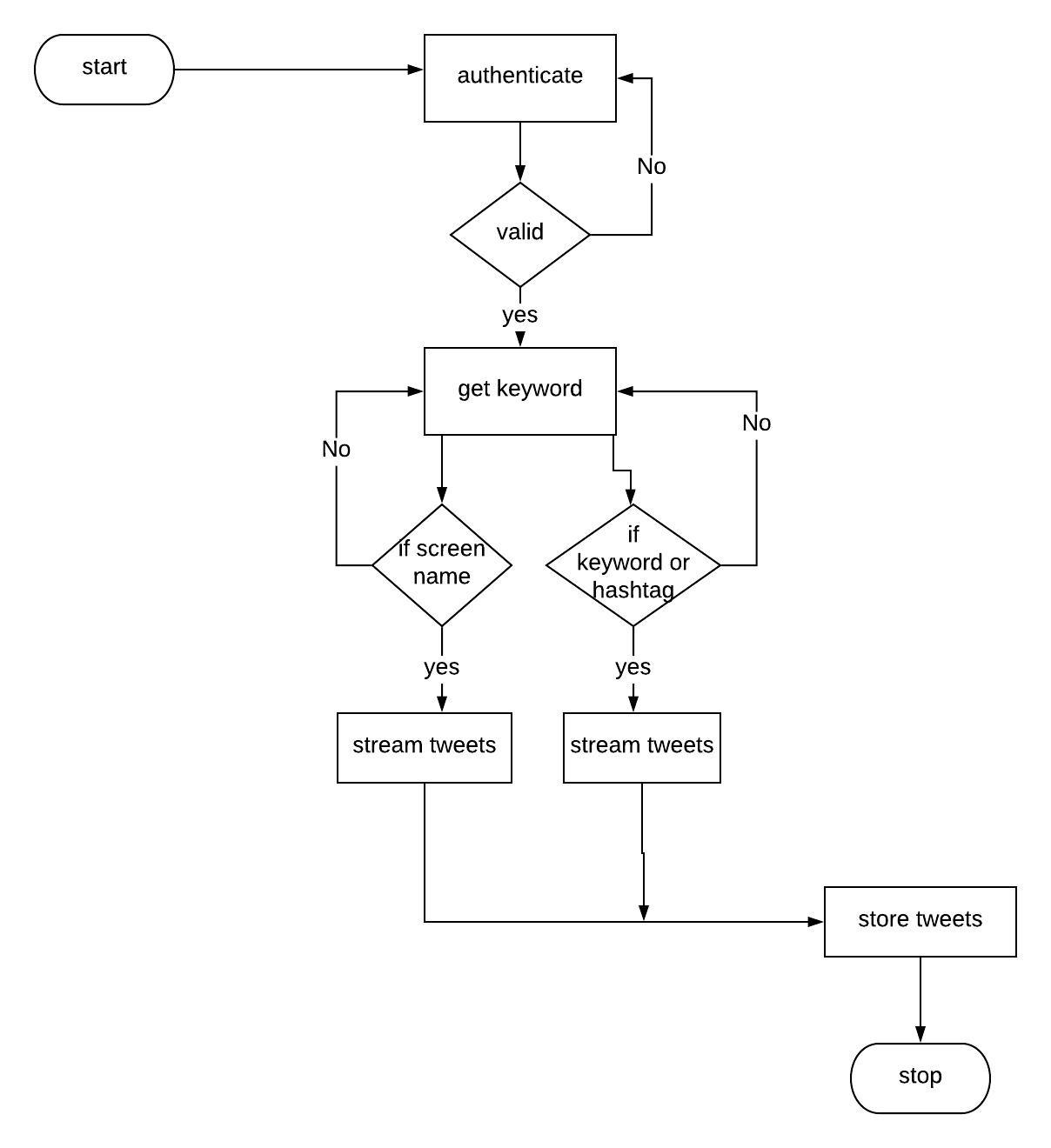
## Flowcharts

It is a graphical representation of a computer program in relation to its sequence of functions (as distinct from the data it processes). In this will show work of our modules.

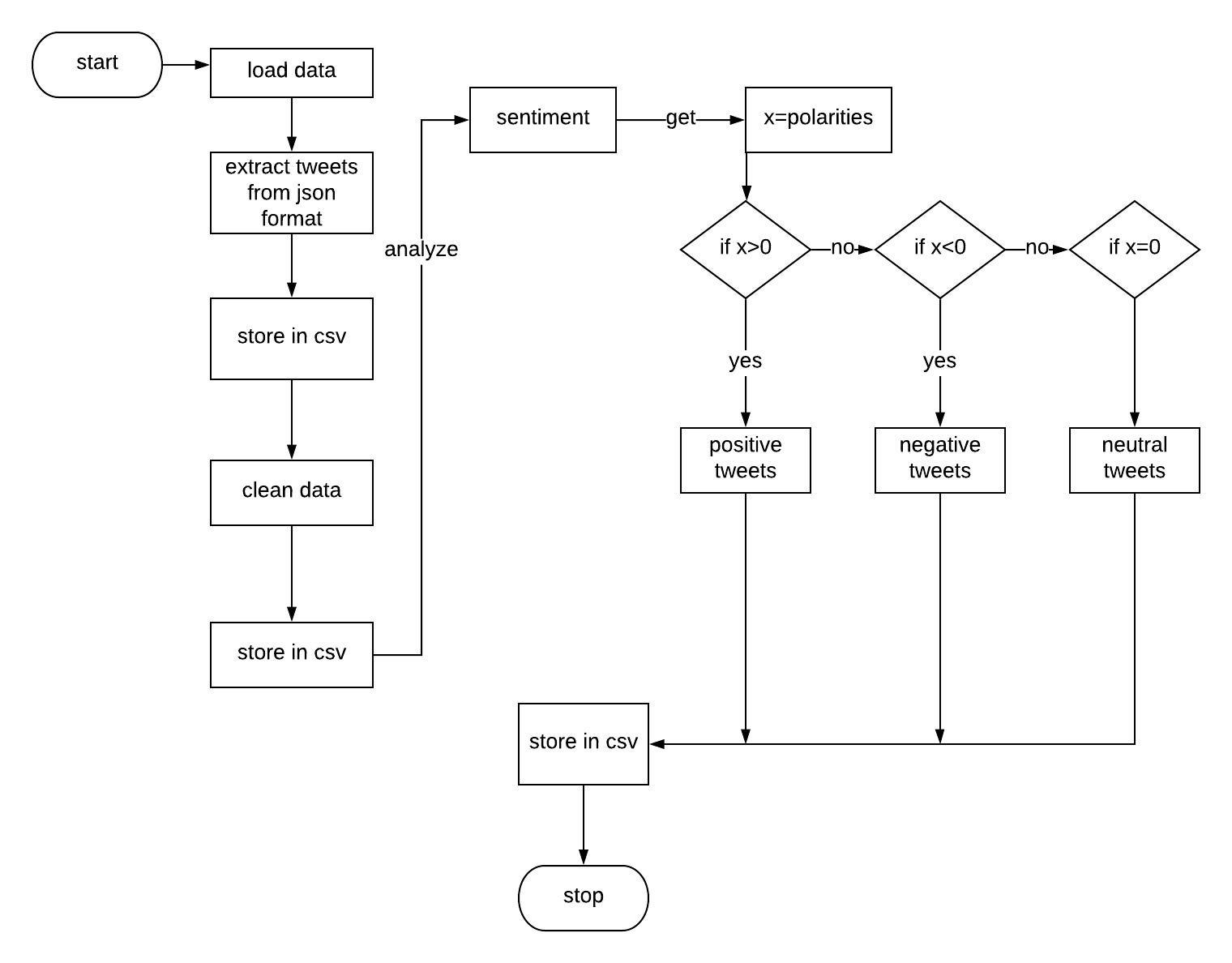
**Flow chart of Overall System**

****

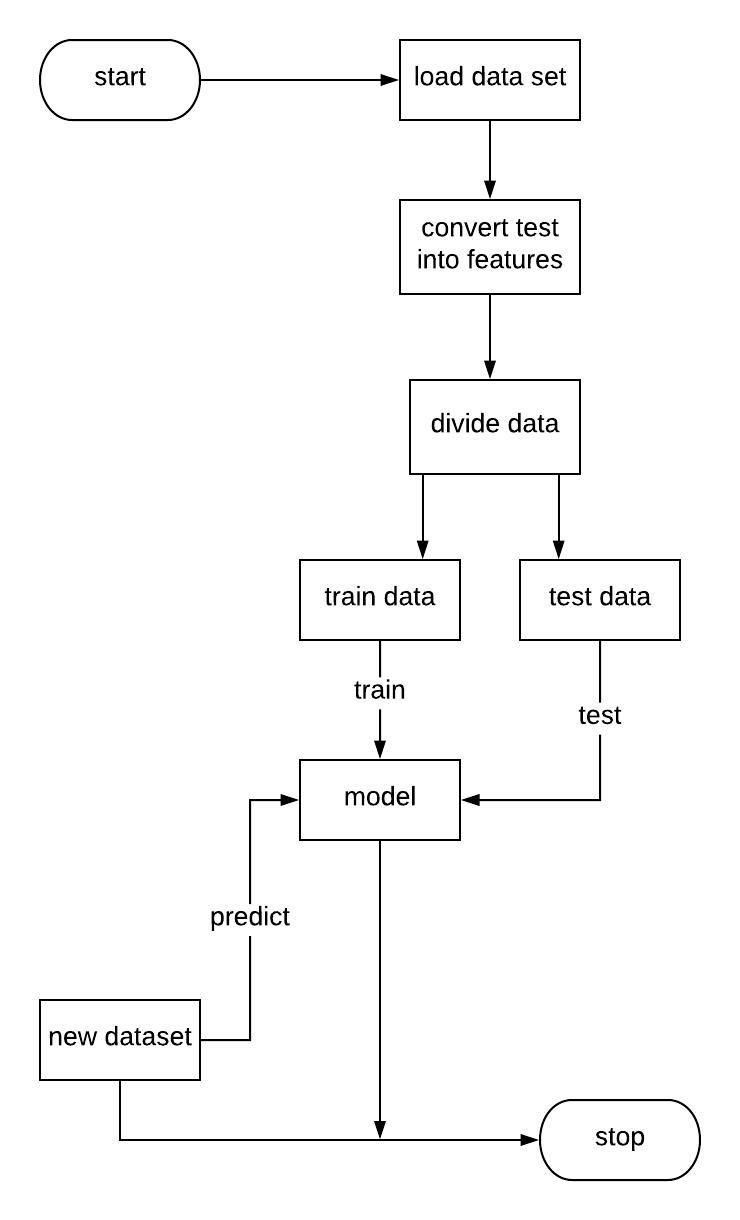
**Flow chart of twitter streaming**



**Flow Chart of Pre-processing and analysis**



**Flow Chart of ML model**



* 1. **Pseudocode**

**6.4.1 Pseudo Code for Twitter Streaming Module**

**Input:** Fetch twitter credentials and tag list, filename

**Output:** Authenticated to twitter and tweet Streaming starts

* If Authentication is True twitter returns the tweets based on tag list in JSON format
* Returned tweets are collected in text file

**6.4.2 Pseudo Code for text extraction from tweets**

**Input:** Fetch the text file of tweet data

**Output:** Returns the csv file containing text part of tweet

* If fetching file is true and data is in json format
* Load the json data and store the text part of tweet into csv file

**6.4.3 Pseudo Code for Pre-Processing the tweets**

**Input:** Fetch the csv file of tweets data

**Output:** Returns the cleaned tweets in a csv file

* If reading csv is true, preprocessing function is called
* When called preprocessing performs removal of special character, hashtags, @user handles, stemming, lemmatizing, stop words removal returns cleaned tweets
* Analyses the sentiment using text blob and returns tweet,sentiment

**6.4.4 Pseudo Code for Building Machine Learning Model**

**Input:** Fetch the dataset

**Output:** Returns the pickled model and accuracy of test data

* If reading the data set is true, coverts the text data into features
* Split features int test and train data
* Train the ML model using train data and test model using test data return accuracy
* Pickle the trained model and return the model

# TESTING

Some points would be considered while testing the GUI which we have made. The points are:

* **Functionality Testing:** We checked whether all the links are working properly or not which are connected to the GUI.
* **Check all the links:**
* Test all internal links.
* Test the Visualization links.
* Test the Pre-processing links.
* Test the extraction links.

1. **Datasets testing:** Make sure the dataset could be in correct format like csv, json etc., At most the dataset was unstructured. At first we need to clean the dataset and later we need to test whether the data become structured or not. And also do check whether we can add or delete the data by manually or automatically.
2. **Usability testing:**

* GUI should be very easy to use.
* Whatever we put on the interface should be very clear.
* Check the inputs which we give are perfect to satisfy its purpose.
* It should be consistent enough.

**Content checking:** Content should be logical and easy to understand. Check for spellingerrors. Usage of dark colors annoys the users and should not be used in the GUI theme. You can follow some standard colors that are used for web page and content building. These are the common accepted standards like what I mentioned above about annoying colors, fonts, frames etc. Content should be meaningful. All the anchor text links should be working properly. Images should be placed properly with proper sizes.

1. **Compatibility testing:**

Compatibility of your GUI is very important testing aspect. See which compatibility testing to be executed:

* Browser compatibility
* OS compatibility
* Mobile browsing
* Printing options
* **OS compatibility:** Some functionality in your GUI is that it may not be compatible with all operating systems. All new technologies used in GUI like graphic designs, interface calls like different API’s may not be available in all Operating Systems. Hence test your GUI on different operating systems like Windows, UNIX, MAC, Linux, and Solaris with different OS flavors.

**7.2 Testing the project:**

**7.2.1 Alpha testing:**

Under Alpha testing, our project was distributed among friends to test and share

their experience each other. After a week of testing, the following bugs occurred.

* Facing problems while connecting to twitter API.
* Facing problems while fetching the tweets. Issue has been resolved by updating the software.
* Problems while storing the tweets. We resolved the problem by using csv and json formats.
* Facing problems at the time of installing packages. We defend this issue by using tweepy package.
* Errors occurred at the time of developing a code.
* Not able to clean the data which we get. We defend this issue by using NLTK tools like tokenizing, stemming etc.,
* Tough to perform few functionalities.

**7.2.2 Beta testing:**

After completion of the alpha testing, we have done some improvements and handover to our mentor. Further he demanded certain modules that were:

* We made GUI which is attractive and we put few modules like visualization, data pre-processing, extraction.
* And we linked all the above to the GUI.
* We also done adding and removing of slang words.
* We also used machine learning algorithms for the accuracy of the tweets.

**8. IMPLEMENTATION**

**8.1 Implementation of the Project**

Process whereby “project inputs are converted to project outputs”. May be looked at as:

* Putting in our all planned activities into action
* Putting into practice what was proposed in the project document (i.e. transforming the Project proposal into the actual project.)
* Team Management of the project and executing the work with correct intentions.

Project Implementation phase involves:

* Project Activation
* Project Operation

This means making arrangements to done with the project. So, we met all the requirements for this project. We have gone through the software library packages, machine learning algorithms, NLTK tools which we are going to use in this project. Accordingly, the language or platform we decided for the project would be simple and robust platform.

**8.2 Project Activation**

### Project activation simply means making arrangements to have the project started. It is the preliminary or foundation stage of project implementation and it involves the coordination and allocation of resources (funds, labor and materials) to make the project operational. The following activities are undertaken during project activation.

* Establishment of Project Implementation Units
* Allocation of the work for team
* Tendering and procurement of software, its installation and testing the platform.
* Preparation of Implantation plans and how to make relevant.
* Designation of the internal project assignments for decision making

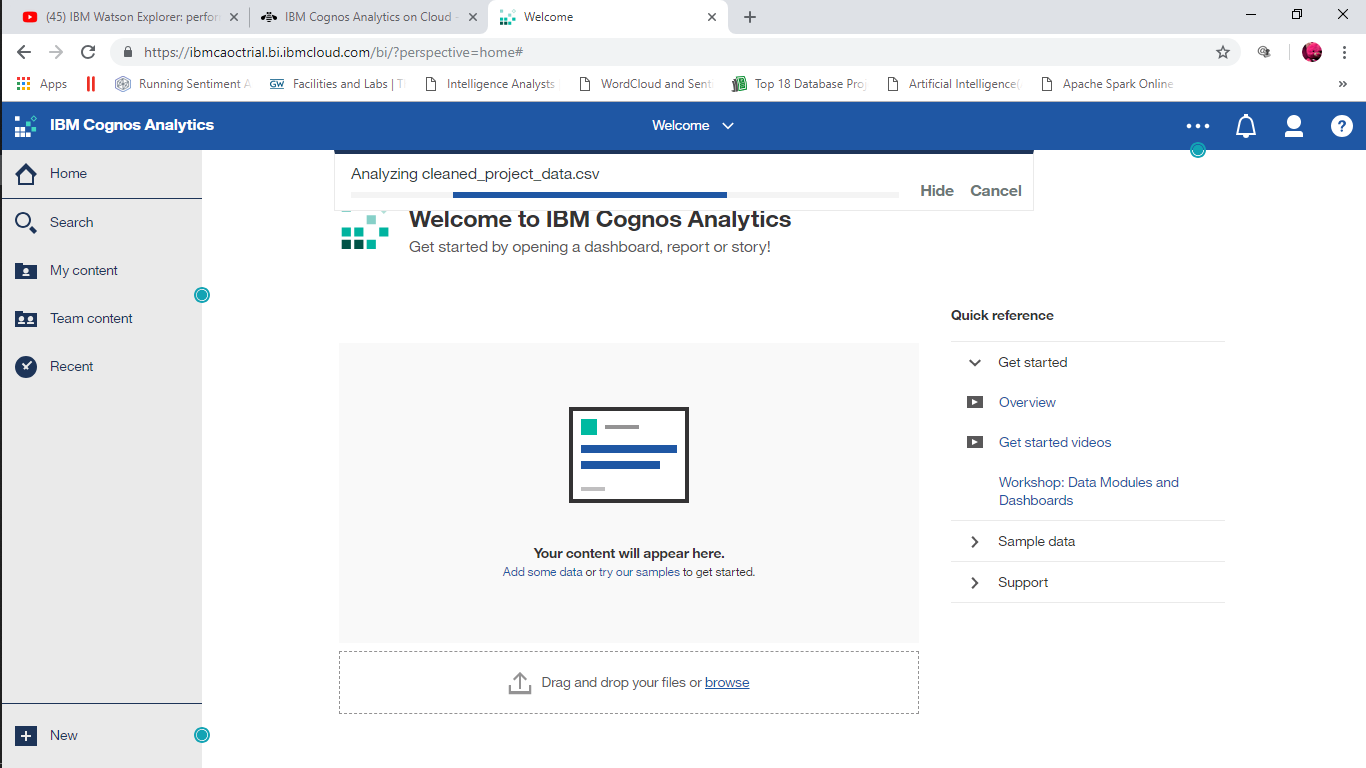
**8.2.1** **IBM CONGO ANALYTICS AND IBM WATSON ANALYTICS**

IBM is a software and services company that offers a wide range of capabilities that enable organizations to make data-driven decisions by applying analytics to information already available. IBM Watson Analytics is an intelligent cloud analysis and data visualization service that allows anyone to quickly discover the motives and meanings of their data, all on their own.

IBM Congo Analytics:

Congo is the business intelligence (BI) and performance management software package IBM. The software is designed to allow commercial users without technical knowledge to extract business data, analyze it, and create reports.

* + 1. **Welcome Portal:**



* Search content in the content of the team, in my content or in recent lists; Open reports, boards, stories and other articles; download files, check your notifications, configure your preferences and home page, and review your schedules and subscriptions.
* Delete, copy, move, edit or execute entries such as reports.
* Create report views for reports without opening the Reports component.
* Create shortcuts to reports, control panels, stories or data modules.
* Establish access permissions.
* We can start creating new content, such as reports, panels or stories, by touching New Report.
* Create and edit a wide range of professional reports. Use templates or customize your reports with directions, bursts, advanced graphics and visualizations.
* Access the IBM Congo’s Analytics Reporting user interface from the Congo Analytics portal by opening an existing report from the computer content or My Content, or by touching New and then clicking on Report.
* For more information, see the Congo’s Analytics Reporting Guide.

.

* + 1. **Reporting**
* Create and edit a wide range of professional reports. Use templates or customize your reports with instructions, bursts, advanced graphics and visualizations.
* Enter the IBM Congo’s Analytics Reports user interface from the Congo’s Analytics portal by opening an existing report from the computer content or My Content, or by touching New, and then pressing Report.
* For more information, see the Congo’s Analytics Reporting Guide.
  + 1. **Dashboards and stories**
* View, monitor and communicate data analysis and analysis with panels or stories. You can assemble a view that contains visualizations such as charts, graphs, tables, tables, maps or any other visual representation of data. Enrich scorecards and stories with web pages and other content.
* You can access the IBM Congo’s Analytics panel or the history user interface from the Congo’s Analytics portal by opening an existing control panel from Team or My Content, or by touching New, and then clicking on Bookmark or History.
* For more information, see the Congo’s Analytics Dashboards and Stories User's Guide.
  + 1. **RESULTS:**
* The results have been noted below with the description on how it works.
* We analyzed from the data collected in other phases and determined their scores, emotions and finally created the dashboards for the comparison.
* The sentiments have been classified with respect to the tweets collected from twitter application.

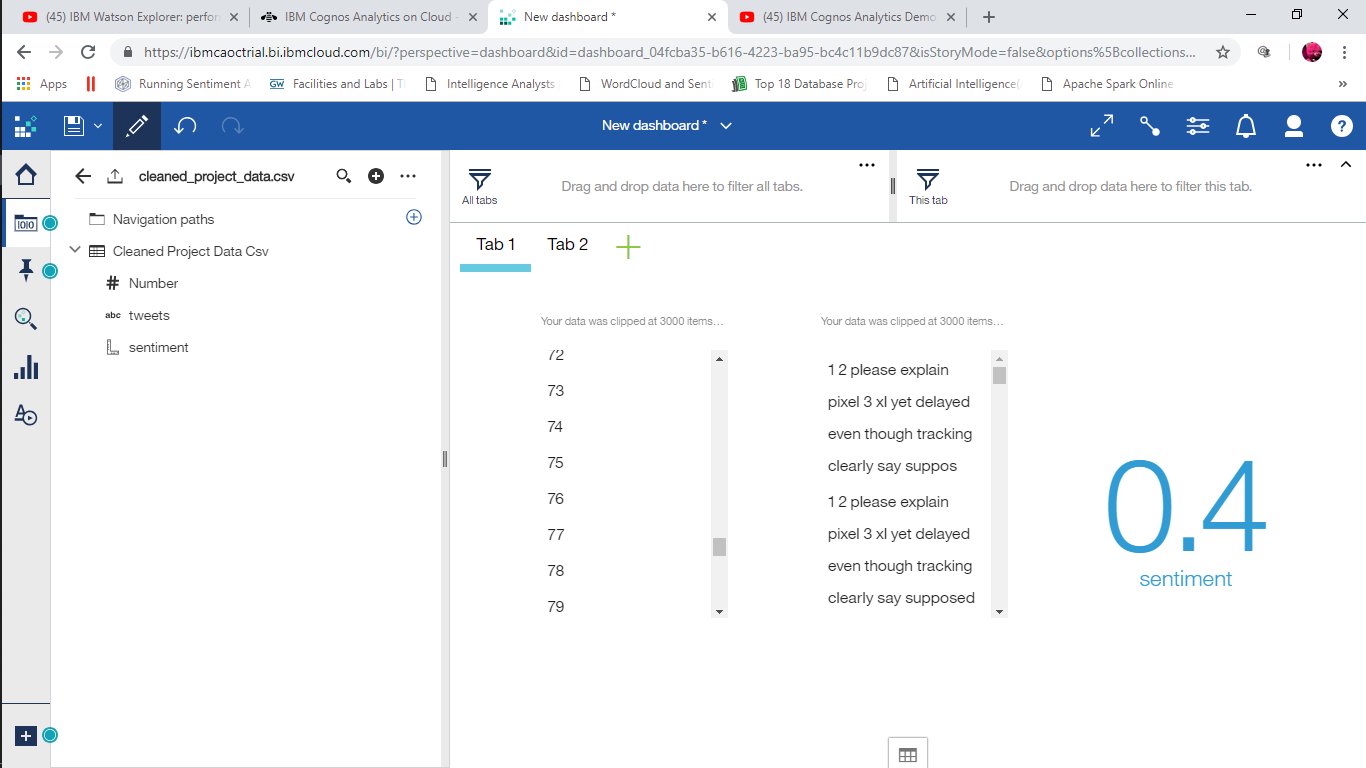


Fig 8.2.1.0: Sentiment score based on tweets on the dataset.

The visualization is all about telling the range of sentiment value attached to each tweets, could be positive, negative or neutral.

* Positive Tweets varies from 0.1 to 1.0
* Negative Tweets varies from -0.1 to -1.0
* Neutral Tweets is 0

Difference in range shows variation in person’s emotions from how positive to negative the emotion is all about.

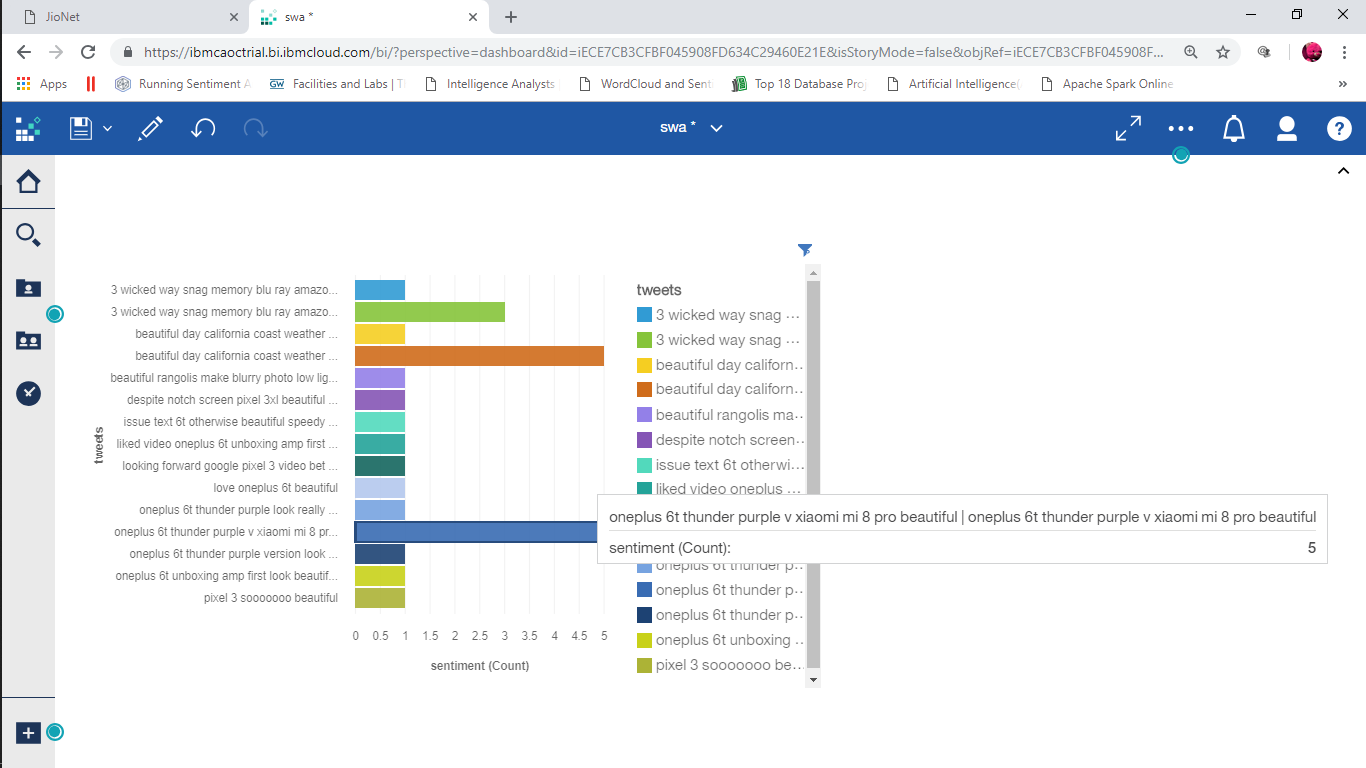


Fig 8.2.1.1: This shows the topic based tweets collected using the keyword “beafutiful”and the re-tweets count.

* The Visualization here tells the user can view how different person react to one single word say” beautiful”. Then it will create a graph using the tweets which has collected for particular tweets.
* The re-tweeted tweets (*i.e.,* some people may comment on the topic again and again) is visulaized with the count.
* Sum of retweetd tweets count :5

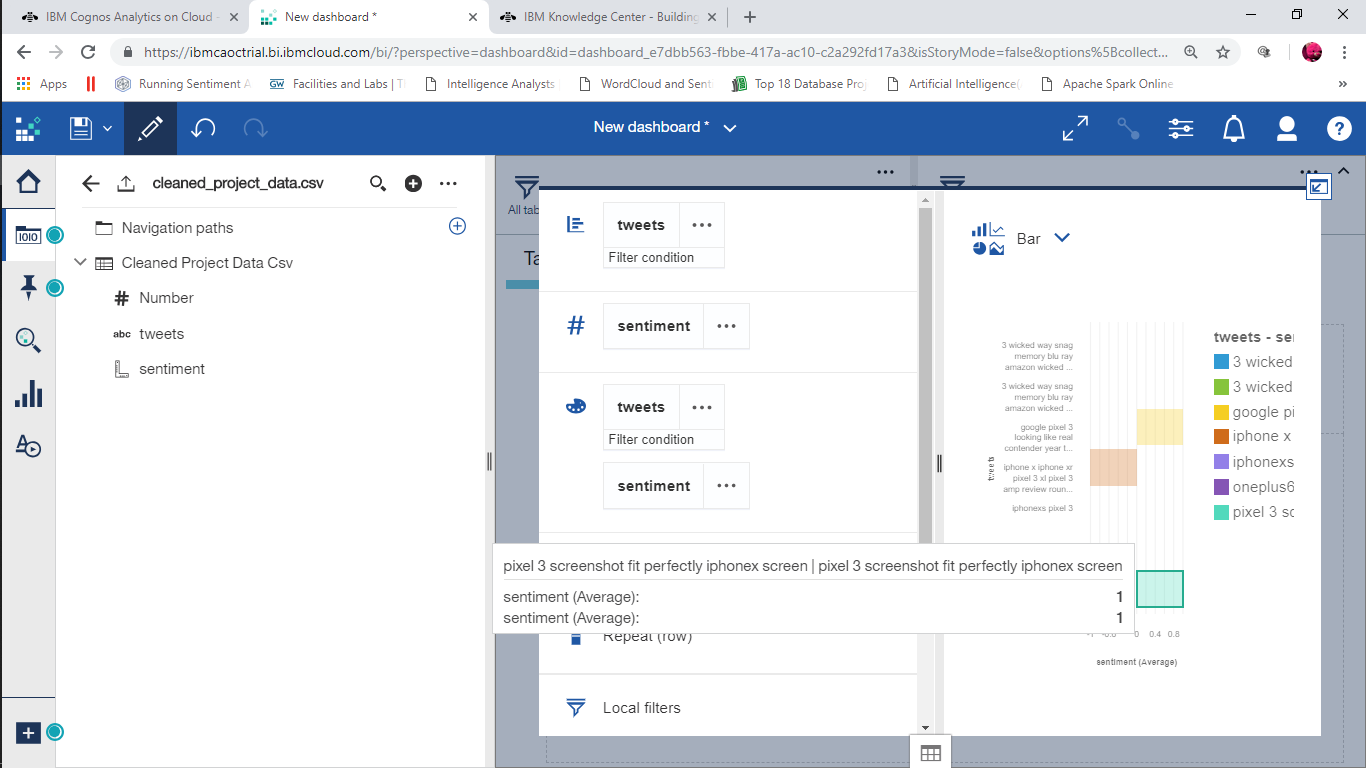


Fig 8.2.1.2 : This shows the analysis on positive sentiment based on tweets and their average.

* The picture describes about how the visualization takes place on the particular topic and checking the emotional or sentimental status regarding that topic. In this we chose iphonex as the topic.
* The positive tweets on iphonex and google pixel 3 is : “green” which is +1.

The below figure Fig 8.2.1.3 depicts :

* This depicts the tweets collected on google pixel 3 as the topic and found the neutral and negative sentiments on that particular topic along with the tweets.
* The Negative sentiments is represented by : “Yellow” which is -1.
* The Neutral sentiments is represented by : “Brown” which is 0.

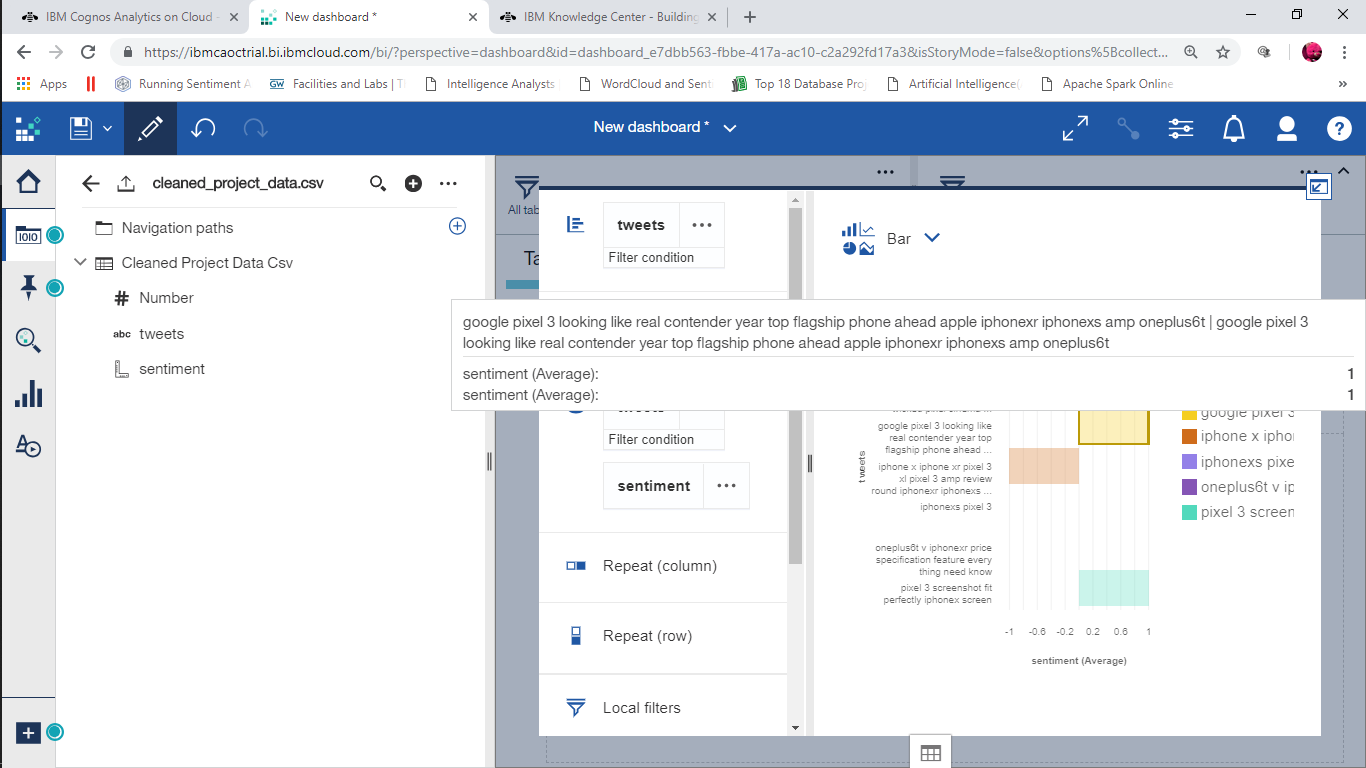
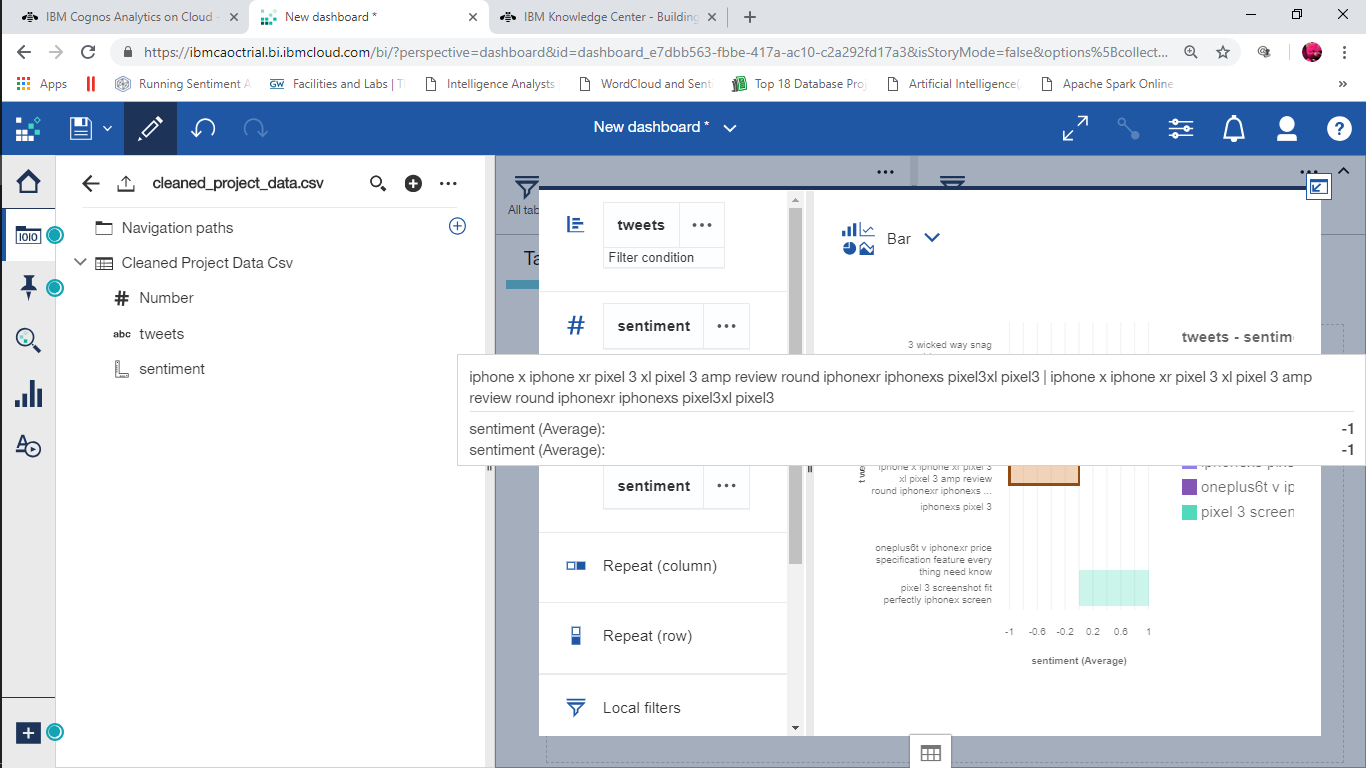


Fig 8.2.1.3.: Shows for the neutral and neagtive sentiments based on the tweets collected.

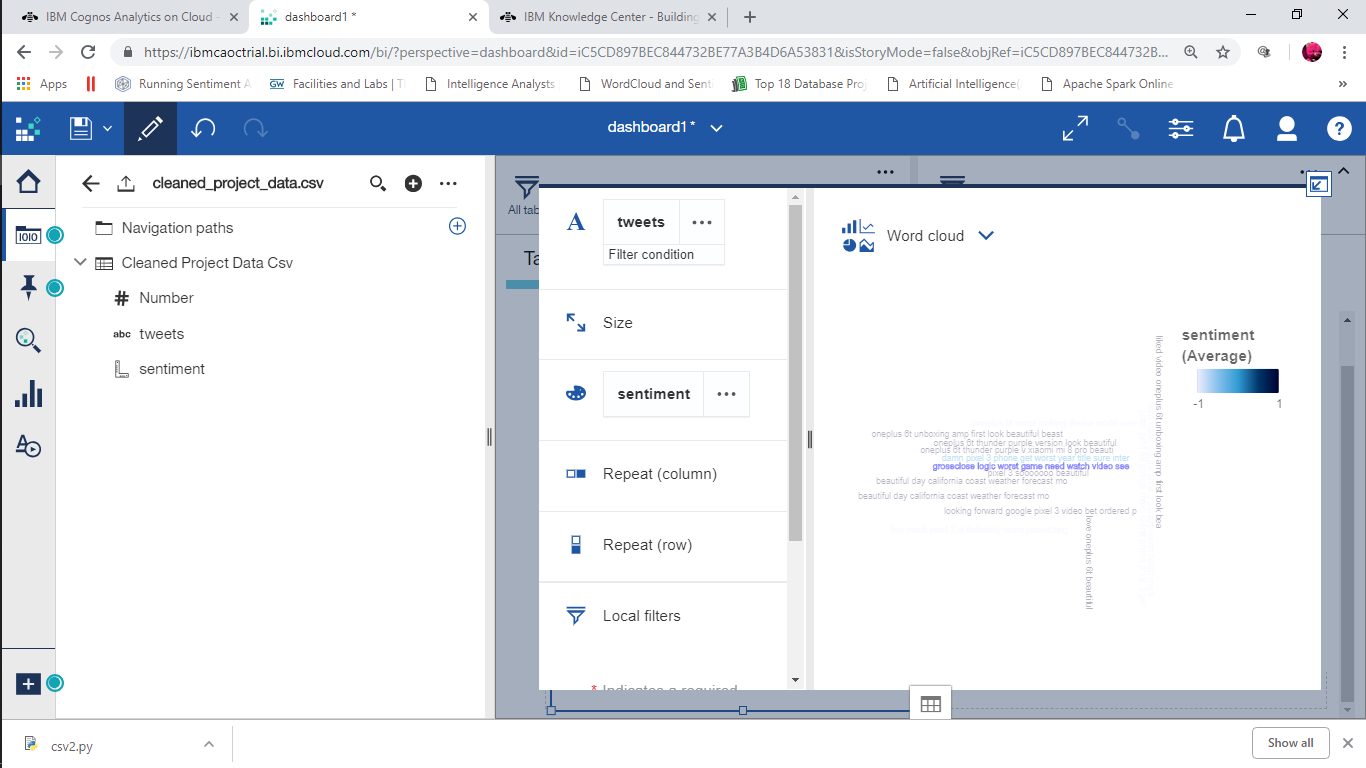
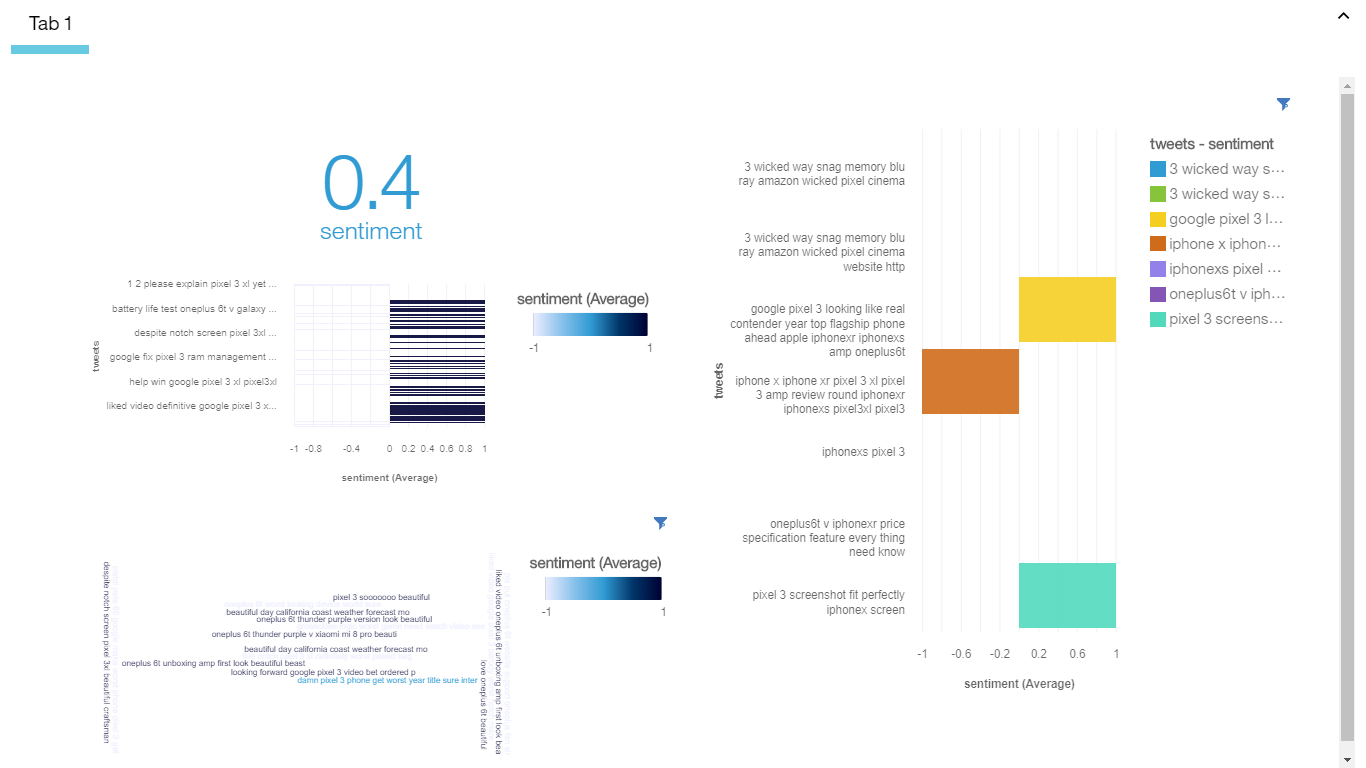


Fig 8.2.1.4: It represents the word cloud based on tweets.

* The word can be cloud created for any particular topic or subject as per the user needful.
* The word cloud canbe created for any topic, word or for any subject.
* The word cloud generated here is based on the topic “ beautiful“.
* The word with negative meaning is given in : “Sky blue”
* The word with positive meaning is given in :” Black”
* The word with neutral meaning is given in : “Blue”

DASHBOARD IN IBM CONGO ANALYTICS

Dashboard consists of all the mini statements of all the results we got off from the collected dataset of report. It consists of wordcloud, sentiment score, bargraph.



**8.2.2 SENTIMENT ANALYSIS USING R**

Nowadays in field of computer science and technology the hottest research area is sentiment analysis. Sentiment analysis can be defined as technique/method of identifying the view of people, given in the form of text regarding to a specific object (event, individual, decision, change etc.). Other synonyms of sentiment analysis are opinion mining, confidence analysis, people attitude towards an object, deriving opinion etc. The main reason behind the popularity of sentiment analysis it gives us overview of wide spread public opinion/thinking related to a topic. Basically, sentiment associated to a particular object is categorized in one of the following category:

1. Positive
2. Negative
3. Neutral

Sentiment analysis is used at multilevel, it can be used to identify sentiment hidden in a document, to be more précised the analysis can be used to calculate the sentiment associated with each paragraph or may be each line. In order to extract the data from twitter we can use Twitter App. Following are steps involved in this-

1. Install and load required packages in R studio to support twitter connection.
2. Create an app on twitter and get the Consumer key (API key) and Secret key (API Secret).
3. Establish the connection between twitter app and R studio with the help of authentication key mentioned in previous step.
4. After successful authentication extract the required data.

**8.2.2.1. Data Collection**

In the data collection process, the data is collected using the live streaming (the data which is already extracted in the other phase of the project). The data is loaded in R Studio using the normal command which is used in R and saving it into other variable for further analysis.

**8.2.2.2. Preprocessing of data for analysis**

Initial phase helps in removal of all other data present in all other languages except the English language. For the next level of preprocessing first we need to load the target file in a user defined object. Further the target data columns are loaded into the data corpus (collection of documents containing (natural language) text. In next step the operations are carried out on the data corpus to clean the data.

**a) Removing punctuations**- English language is supported by different punctuation marks such as dot (.), coma (,) etc. However, punctuations are meaningless whenever we have to perform analysis, so it is become important to remove the punctuation marks.

**b) Removing white spaces -**It might be possible that extracted text contain unwanted whitespaces, which may act as noisy data during analysis. For better results, it is advisable to remove the white spaces.

**c) Converting all the text into lower case-** Most of the analysis/mining code treat are case sensitive, so to reduce errors it is advisable to have our all our text in same case (lower case or upper case.

**d) Removing the stop words of English**- Stop words are comprised of general words which is to support our sentence such as I, me, my, do, should etc. However, these are not important from the data analysis point of view, so it is advisable to remove such words.

**8.2.2.3. Visualization**

The data which extracted in the other phase of theproject is used for the visualization. The column in the dataset sentiment (which consists of the sentiments as 0 for neutral, 1 for positive and -1 for negative sentiments) was in integer form. For analysis the sentiment column was converted into character form and analysed using R. The packages used for the sentiment analysis was:

* dplyr
* syuzhet
* lubridate
* ggplot2
* reshape2
* scales

The graph used for the visualization is barplot. A barplot is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable).

**8.2.2.4 PSEUDOCODE**

Even though the result is attached, the rough pseudocode for the entire analysis is given below:

Step 1: Extraction of data (In this, the data which has already extracted in the other phase of the project is used. No real time extraction)

Step 2: Importing the CSV file to Rstudio for further analysis

Step 3: Building the corpus

Step 4: Clean the text

Step 4.1: Remove punctuations

Step 4.2: Remove stop words

Step 4.3: Conversion of upper case to lower case

Step 5: Import libraries

Step 6: Obtain the sentiment score

Step 7: Plot the bar graph

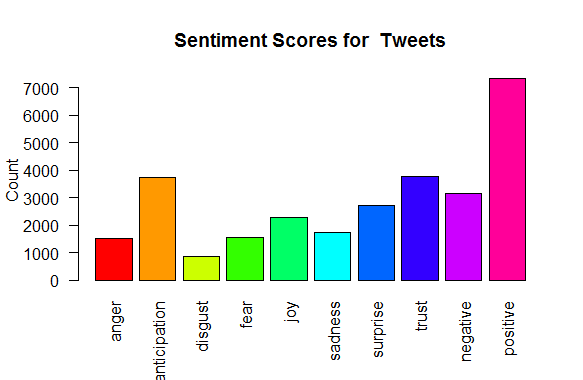


Fig 8.2.2.1: The figure shows the sentiment score for tweets and classified them on basis of counts.

The analysis of the sentiment scores is classified into various sentiments like anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive. Result has been analyzed using the barplot as:

* Anger: 1500
* Anticipation: 3700
* Disgust: 1000
* Fear: 1800
* Joy: 2200
* Sadness: 2000
* Surprise: 3000
* Trust:3500
* Negative: 3100
* Positive:7000

**8.2.3. SENTIMENTAL ANLYSIS USING TABLEAU**

As a leading data visualization tool, Tableau has many unique and desirable features. Its powerful discovery and data mining application lets you answer important questions in seconds. Anyone can use Tableau's drag-and-drop interface to display data, explore different views, and even easily combine multiple databases. It does not require any complex script. Anyone who understands the company's problems can solve it by consulting the relevant data. After the scan, sharing with others is as easy as posting to Tableau Server. Tableau is a business intelligence tool for visually analyzing data. Users can create and distribute an interactive and shared dashboard that represents trends, variations, and data density in the form of tables and tables. Tableau can connect to files, relational sources, and Big Data to acquire and process data. The software allows the combination of data and collaboration in real time, which makes it very unique.

Tableau’s remarkable features was the one of the reasons why we chose this tool for our project.

1. Standalone: ​​Tableau does not need a complex software configuration. The desktop version used by most users is easily installed and contains all the necessary functions to start and complete the data analysis.
2. Visual discovery: the user explores and analyzes data using visual tools such as colors, trend lines, graphics and graphics. There are very few scripts to write because almost everything is done by dragging and dropping.
3. Combination of different data sets: Tableau allows you to combine different relational, semi-structured and unprocessed data sources in real time, without high initial inclusion costs. Users do not need to know the details of how the data is stored.
4. Centralized Data: Tableau Server provides a centralized location for managing all published data sources in the organization. You can delete, edit permissions, add tags, and manage programs in one place. It's easy to schedule extraction updates and manage them on the data server. Administrators can centrally configure an extraction program on the server for a full and incremental update.
5. For the visualization we use pie-chart, bubble chart, pack bubbles and dashboards.

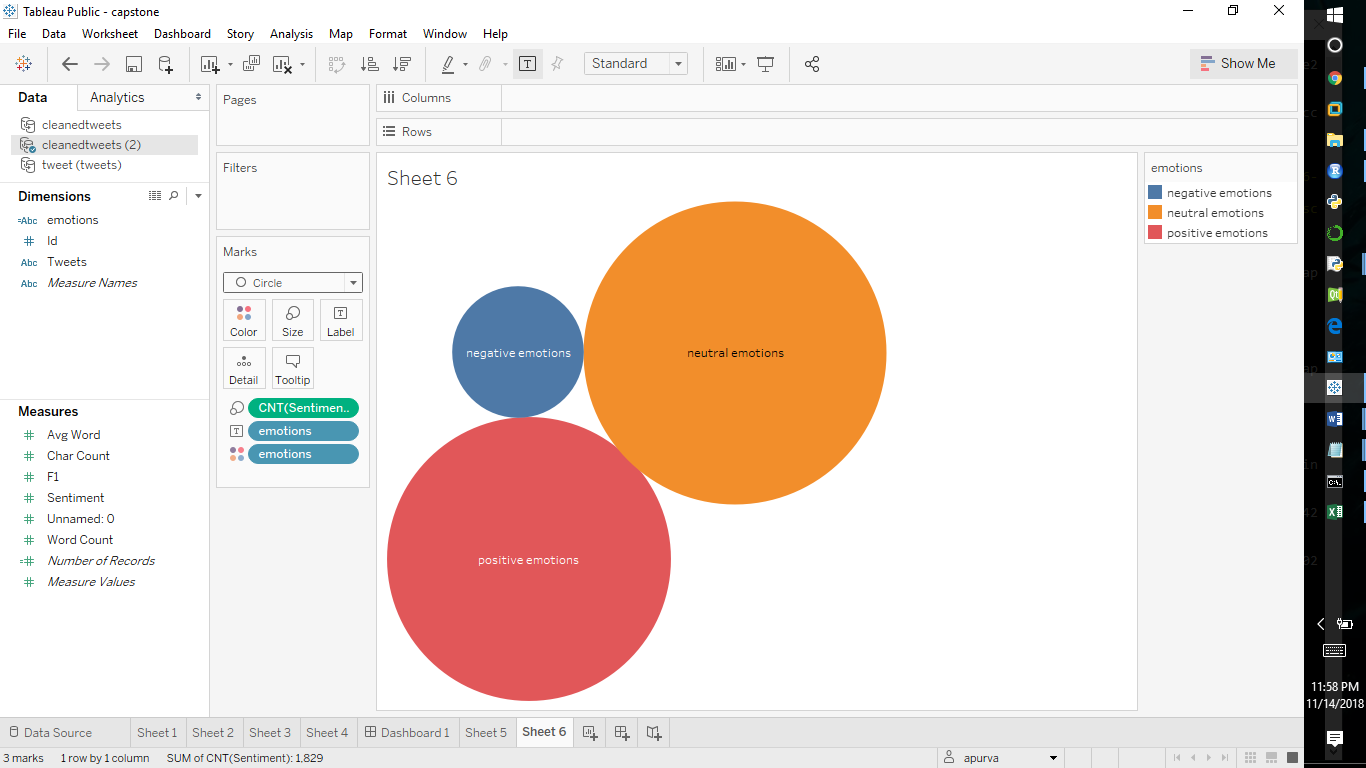


Fig 8.2.3.1: The tweets classified based on the sentiments with assigned scores.

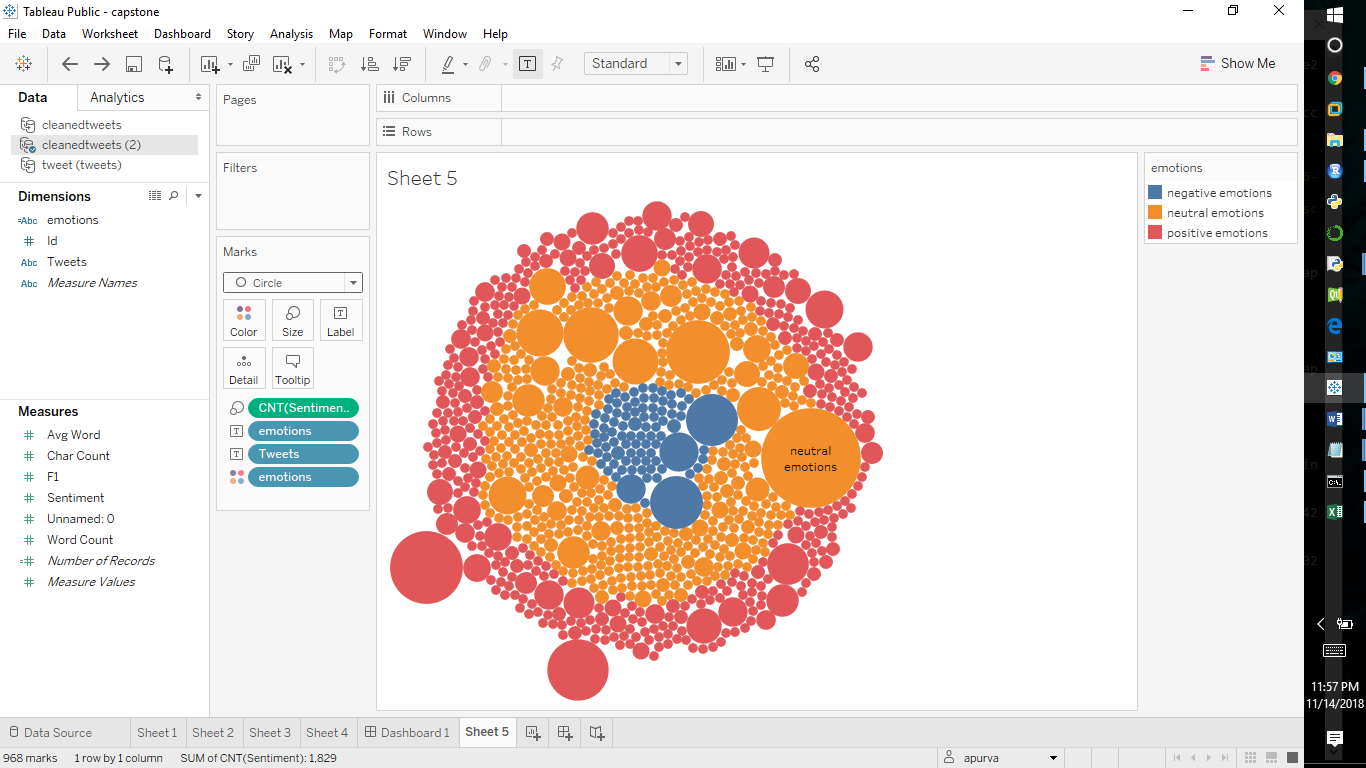


Fig 8.2.3.2. It is the modified version of Fig1. Fig 2 says that the negative comments are clustered and brought into the centre of the graph. While the neutral tweets were centered after the negative tweets. They are surrounded by the positive tweets which is shown in red colour.

* The classified positive sentiments based on tweets: “red”, +1 is assigned.
* The classified negative sentiments based on tweets: “blue”, -1 is assigned.
* The classified neutral sentiments based on tweets: “orange”, 0 is assigned.

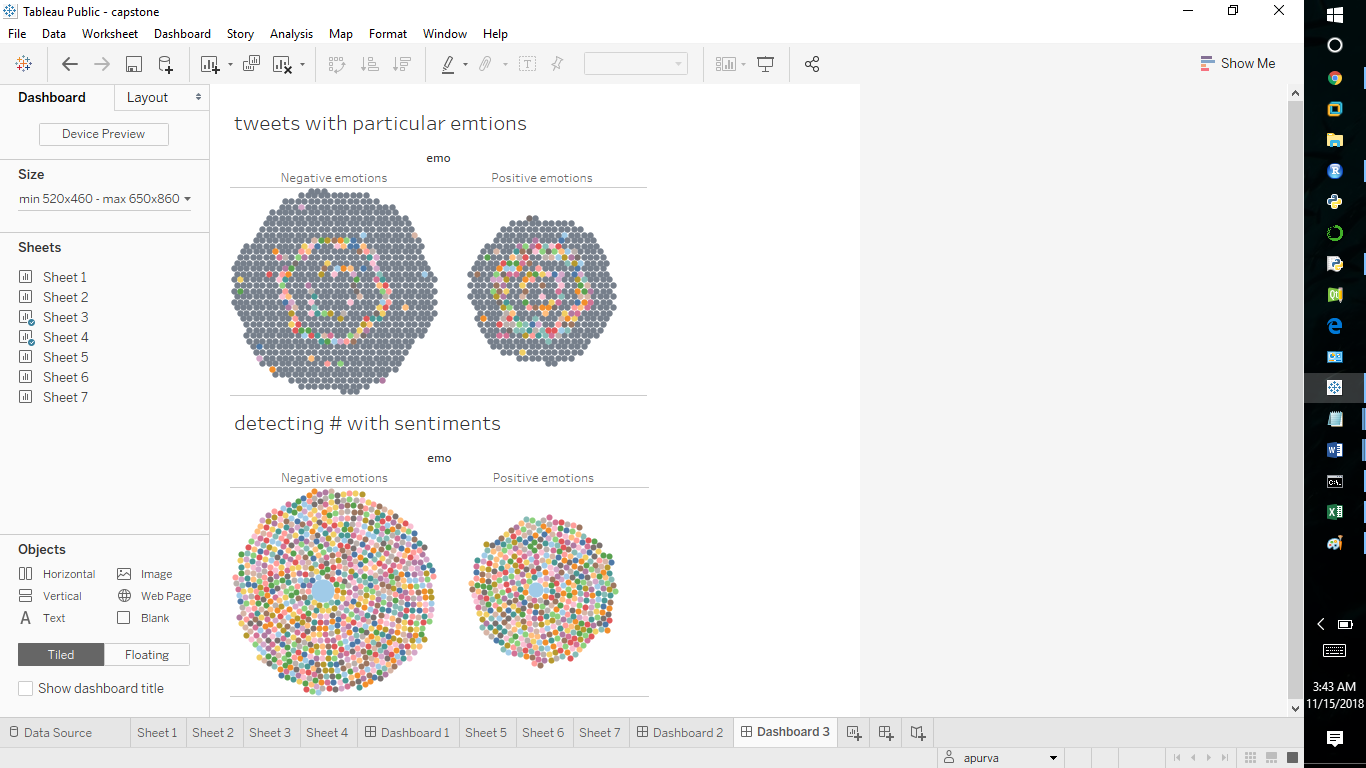


Fig 8.2.3.3: detecting hashtag and @.

* This colorful portion above is the hashtag and other portion is the normal tweets.
* In second figure, it explains as:

Centre portion defines: hashtag in blue area

Rest portion defines: normal tweets

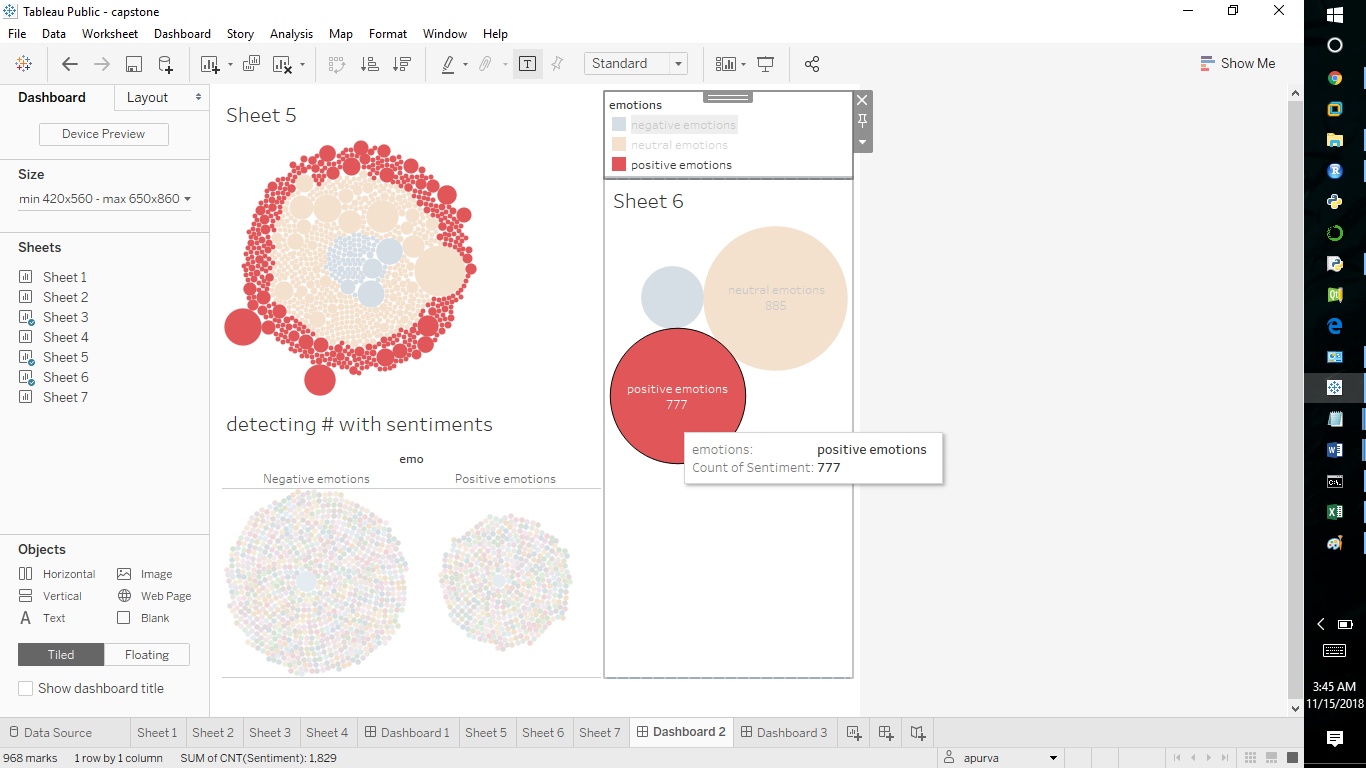


Fig 8.2.3.4: This shows the dashboard for positive and negative emotions.

* While clicking on a particular area i.e., in this the red area it will show for the positive tweets.
* The count for positive tweets are: 777

**8.2.4. VISUALIZATION OF SENTIMENTAL ANALYSIS USING PYTHON**

Data visualization usually describes any effort to help people understand the meaning of the data by placing it in a visual context. The patterns, trends and correlations that may not be detected in the textual data can be exposed and recognized more easily with the data visualization software. For the visualization, we have used ggplot2, matplot lib. For making the visualization more effective we have created a GUI using the pyQt5. GUI is not so complicated, with the help of import csv button it will import any csv files. Further details of the GUI will be explaining in the upcoming chapters.

Following are the libraries used:

* ggplot2: ggplot2 is a graphing system for Python based on ggplot2 of R and graph grammar. It is designed to achieve a professional appearance, quickly trace with a minimum code.
* matplotlib: matplotlib is a tracking library for the Python programming language and its numeric extension NumPy. Provides an object-oriented API to embed graphics in applications using general GUI tools such as Tkinter, wxPython, Qt or GTK +.
  + - 1. **Creation of GUI using Python**

The GUI is created in python with the following option that provides to the user for the interaction. The GUI is user friendly and makes its user interactive by automatically loading the entities and creates the bar plot.

The GUI in front end consists of:

* PreProcessing
* Visualization
* Login
* Exit

Visualization: It is connected to the another page. As soon as the user clicks on the button for visualization then it will redirect to the page that will show for importing csv and plot button. The import csv will help in uploading the file of csv which is the cleaned data used in the previous phase of the project. And the other button i.e., plot button will plot the trend lines, barplots of sentiments etc.

The GUI has the following features:

* The GUI is created using python with pre-processing and visualization buttons.
* Along with this there are two more buttons login and exit.
* If in case the login is not working, then exit from GUI using exit button.
* The visualization will work automatically as clicking on the buttons.
* If we click on pre-processing the tweets collected from live using the twitter application

and the processing it then it will pre-process the tweets.

* After the pre-processing of tweets, it will visualize the data which is the collected tweets.

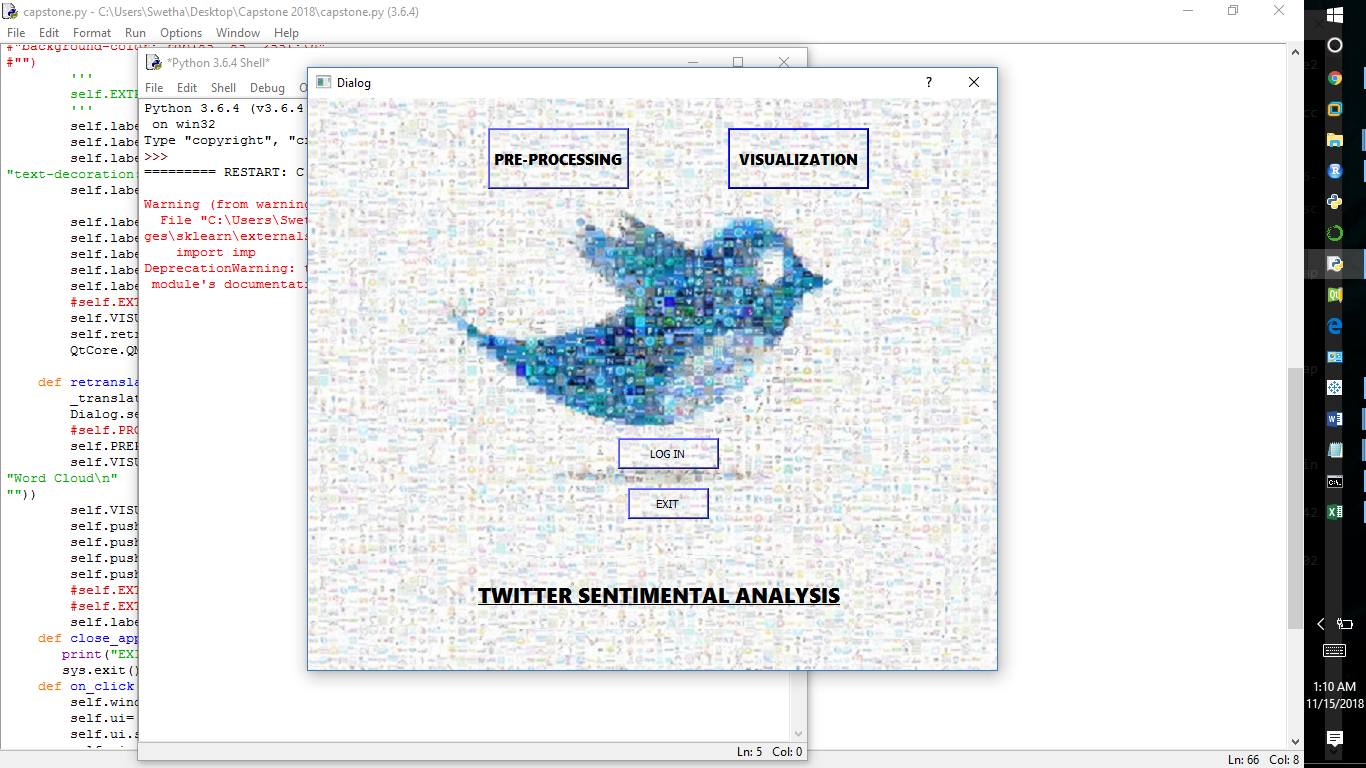


Fig 8.2.4.1.0: GUI created using python.

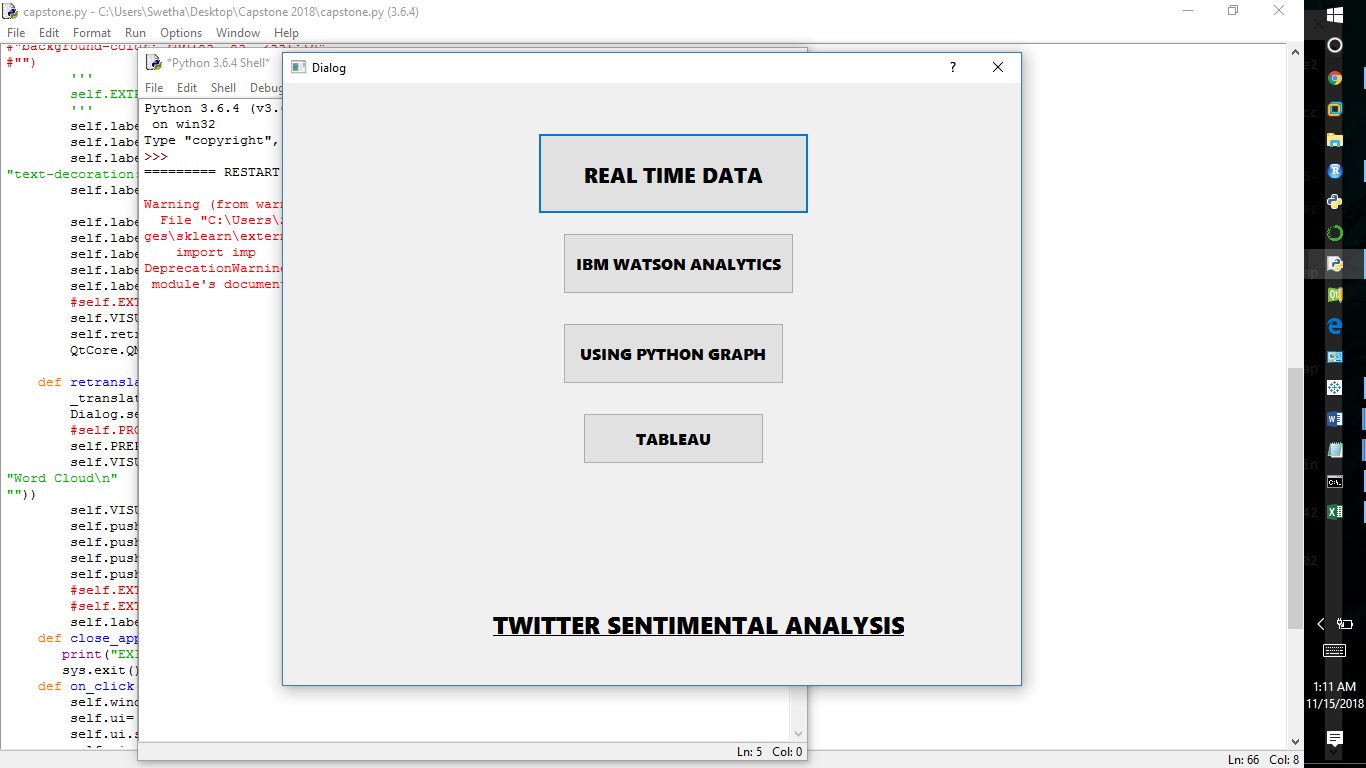


Fig 8.2.4.1.1: This shows the second window of the GUI.

* The GUI has real time data for taking the real time data.
* IBM Watson Analytics will connect for Watson Analytics.
* Using Python graph it will show the another window with two options.
* Tableau will connect to tableau sever for further analysis.

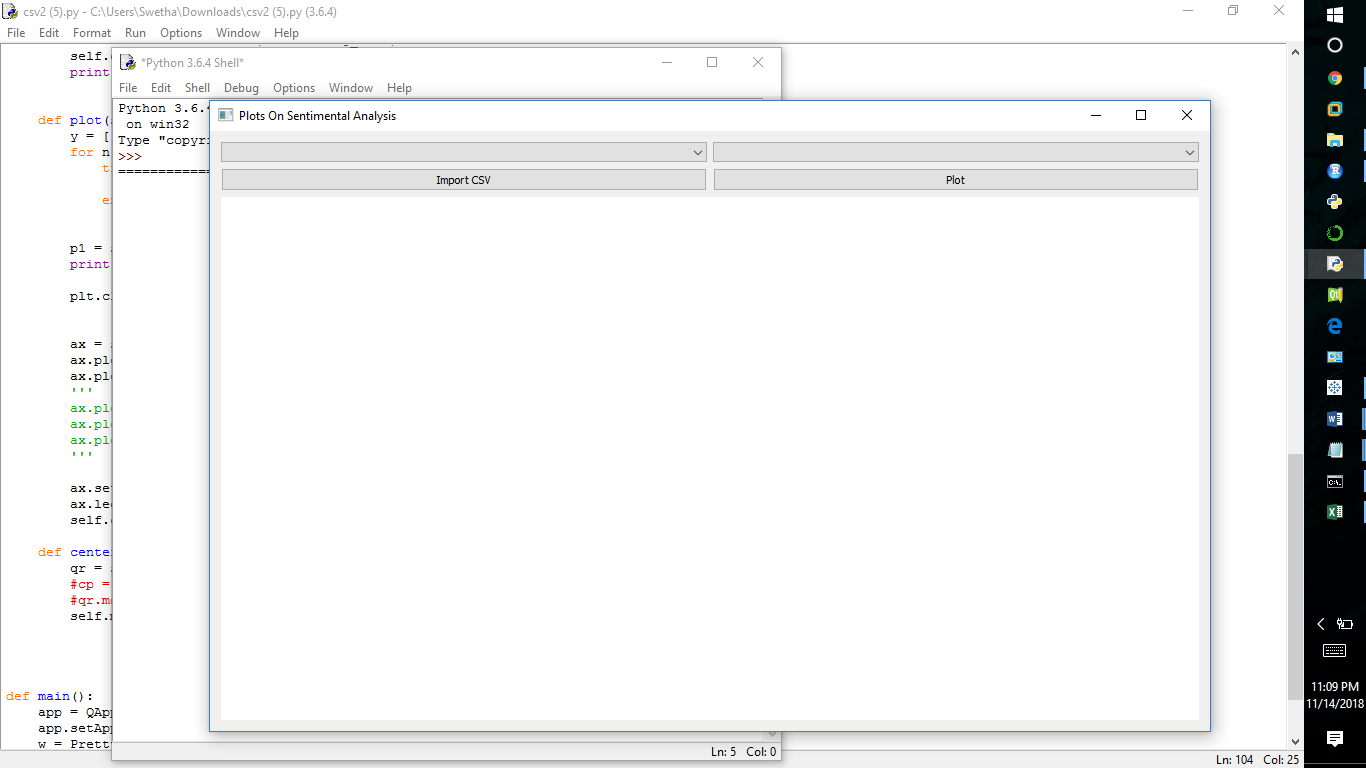


Fig 8.2.4.1.2: It shows the sample of importing a csv file.

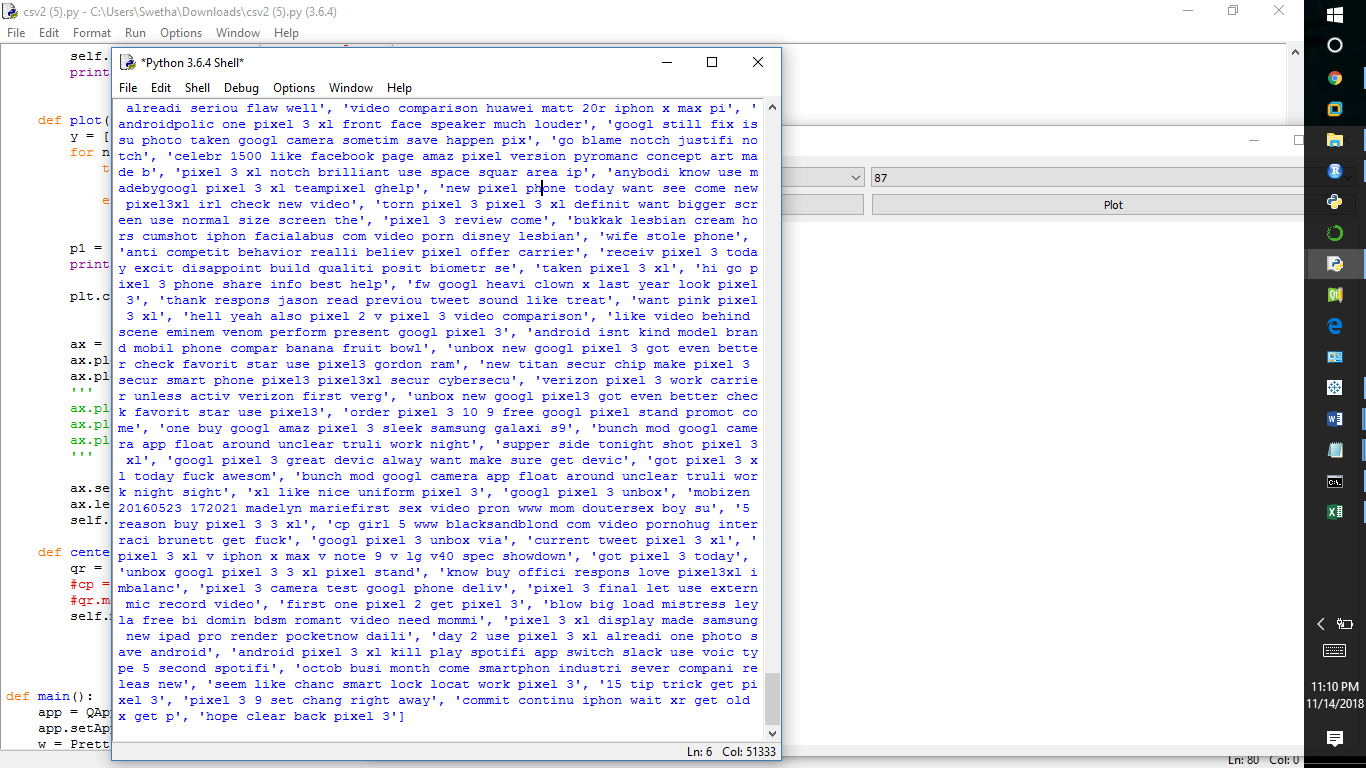
* It is showing how the GUI is looking on while clicking the visualization button from the main GUI.
* As soon as the user clicks on import csv, it will add particular csv file that user wants to upload. 

Fig 8.4.2.1.3: The is the snapshot of tweets collected using python.

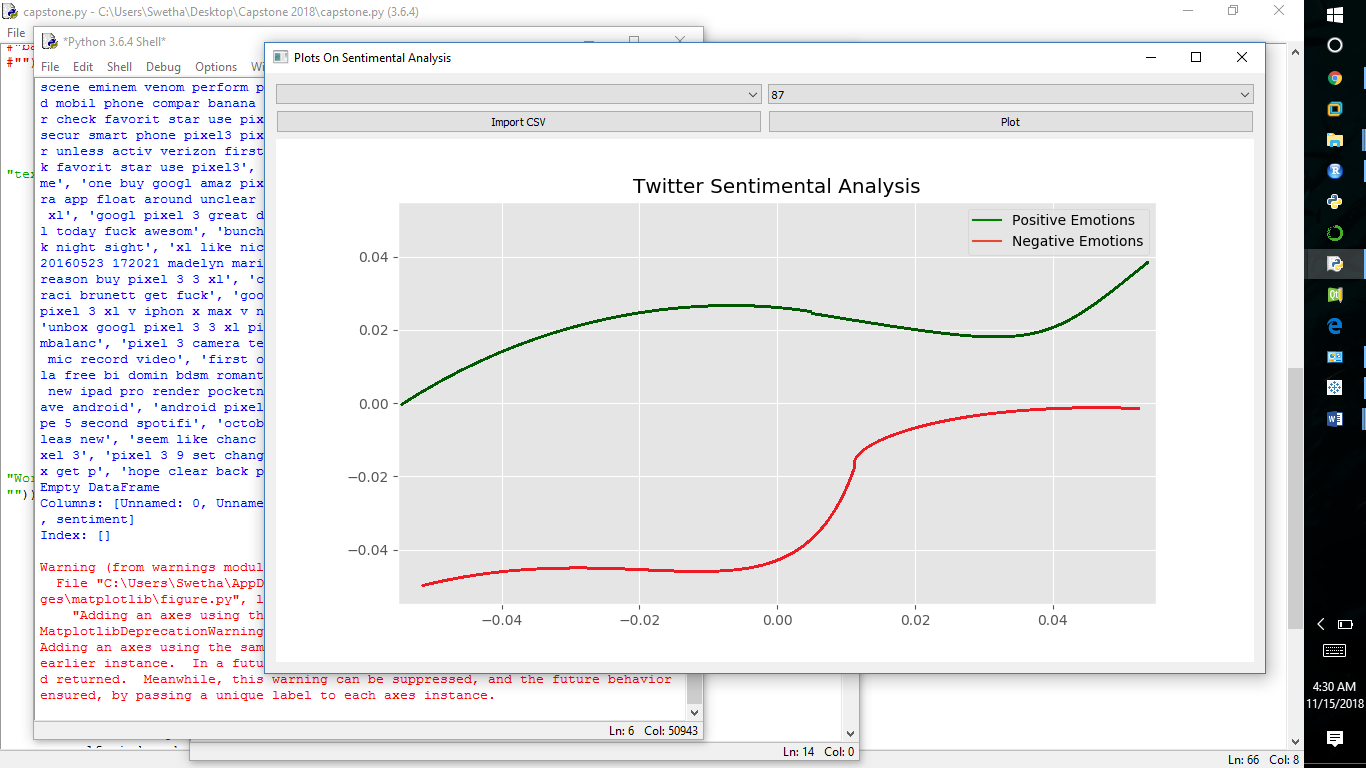


Fig 8.4.2.1.4. this is the snapshot of positive and negative emotions.

* The positive emotions are represented in: “green”
* The negative emotions are represented in: “green”

The reason behind of using different tools in sentimental analysis is described below in form of table.

|  |  |  |  |
| --- | --- | --- | --- |
| SR. NO: | TOOLS USED | ADVANTAGES | DISADVANTAGES |
| 1. | IBM Watson Analytics or IBM Congo Analytics | * Helps in real time analytics by directly connecting with twitter. * Easy to handle | * Poor internet connection may loss the work done so far. * Need internet connection . |
| 2. | Using R STUDIO | * Great features for data visualization. * Endless libraries * Ad-hoc analysis can be done better. | * Programming skill required * Non-statisticians can’t use r easily. |
| 3. | TABLEAU | * More interactive and handles large amount of data. * Can create dashboards * Can connect to external server using local host. | * Scalability * Embeddability * No custom visual imports * Static an single value parameters. |
| 4. | PYTHON | * Better for data manipulation * Easier for understanding. * Python can work easily | * Not much user friendly. * Code complexity is higher. |

Table 1: It shows the comparison of different tools used in sentimental analysis.

**8.3 Project Operation**

This is practical management of a project. Here, project inputs are transformed into outputs to achieve immediate objectives. In this the basic requirement we have taken from twitter i.e., at first we need to make an account in apps.twitter.com. From this website twitter would give permission to access the tweets for our analysis purpose like to analyze positive, negative and neutral tweets. In apps.twitter.com twitter will generate consumer key, consumer secret key, access token, access token secret to access the twitter and fetch the live tweets.

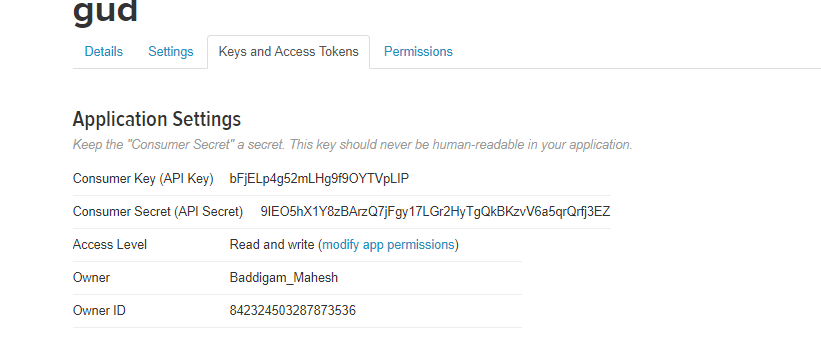


Fig 8.2.0 Setting the twitter application



Fig 8.2.1 Accessing the tokens from twitter application

**8.3.0** **COMPARISON OF** **SENTIMENTAL ANALYSIS USING SPARK IN R VS PYTHON**

Sparklyr is an R package that allows you to analyze data in Spark while using familiar tools in R. Sparklyr supports a full backend for dplyr, a popular tool for working with data frame objects both in memory and out of memory. You can use dplyr to translate the R code into Spark SQL. Sparklyr also supports MLlib so you can run classifiers, regressions, clusters, decision trees and many more machine learning algorithms in your data distributed in Spark. With sparklyr you can analyze large amounts of data that traditionally would not fit in the R memory. Then, you can collect the Spark results in R for further visualization and documentation. Sparklyr is also extensible. You can create R packages that depend on sparklyr to call the complete Spark API.

**8.3.0.1. Data Import**

Another objective of this article is to encourage the reader to try it. A simple Spark local mode session is used. The data is imported from the local machine in which we have already saved the data extracted and cleaned from twitter for sentimental analysis.

**8.3.0.2. Spark Connection**

We are using spark\_connection for the locally using spark\_connect (master=” local”). After that it will automatically connect to spark.

copy\_to function of dplyr library to copy the content i.e., the dataset to spark connection.

**8.3.0.3. Spark mlib:**

Spark has an automatic learning library. And we use the Naïve Bayes algorithm to predict the accuracy. Naive Bayes classifiers. Supports Multinomial NB (see here), which can handle discrete data with finite support. For example, when converting documents into TF-IDF vectors, it can be used for document classification. By making each vector a binary data (0/1), it can also be used as Bernoulli NB. The values ​​of the input characteristic must be non-negative.

**8.3.0.4. Pseudocode**

1. Connecting r to spark using sparklyr.
2. Connecting it through the function of spark\_connect (master=” local”)
3. Loading the data sets
4. Copying the dataset into spark connection.
5. Creating the testing and training datasets on different probability ratios.
6. Using “mlib”, ml\_naive\_bayes, ml\_logistic\_regression, ml\_random\_forset,

ml\_kmeans the function for multi-nomial Naïve Bayes, Logistic Regression

, Random Forest, K-Means Clustering using Spark in R.

1. Predicting the accuracy of the sentiments.

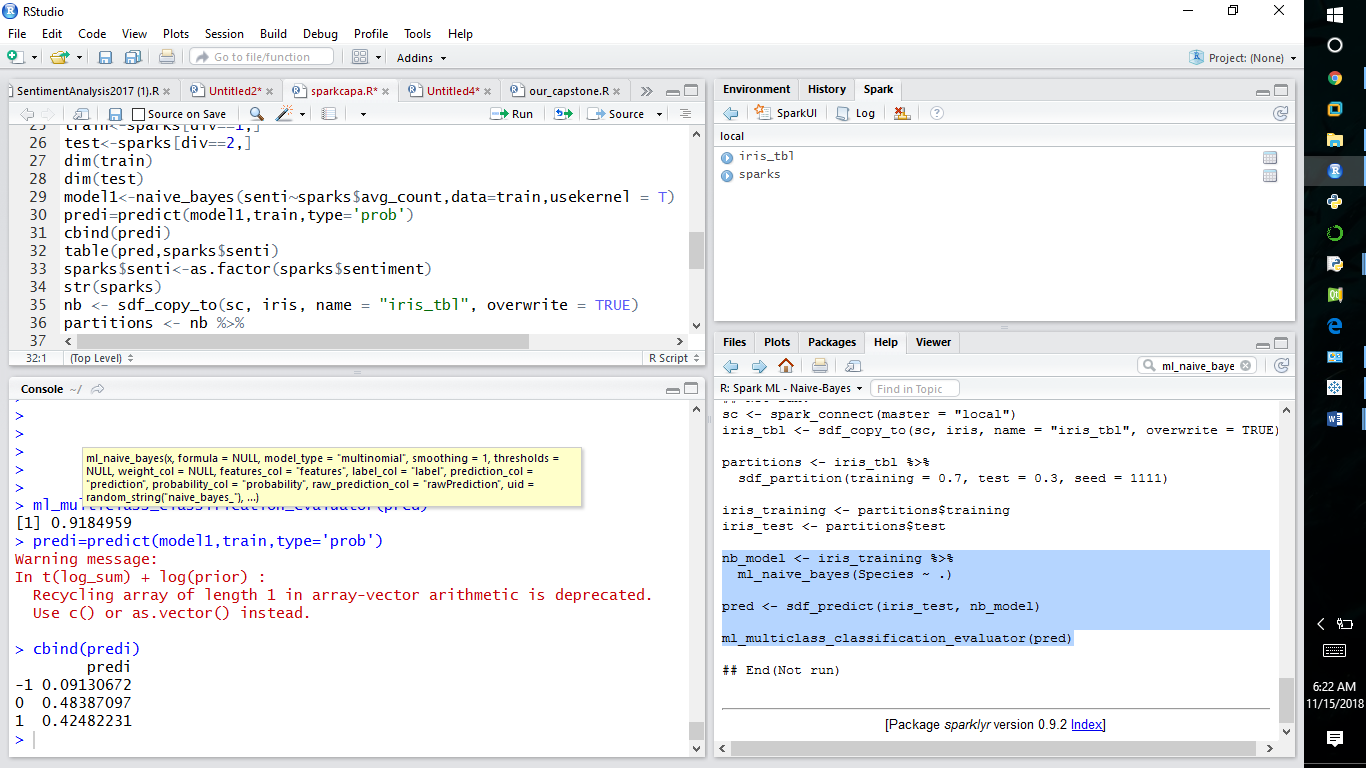
**8.3.0.5. Result in spark R**

Fig: shows the accuracy on sentiments using spark in R

* The probability of positive emotions :0.424
* The probability of negative emotions :0.091
* The probability of neutral emotions :0.483
* Overall accuracy: 91%
* Similarly the accuracy for other algorithm are calculated

**Comparison of sparkr over python**

|  |  |  |
| --- | --- | --- |
| FEATURES | PYTHON | SPARKR |
| Speed | * Python speed is lesser as compared can be | * It is 100 times faster than accessing the data from Hadoop and 10 times than accessing data from disk. |
| Computation | * Computation is slower and doesn’t have in memory computation. | * In memory computation |
| Language complexity | * It is easy as it is a general language. * Easier to understand than Scala. | * It is short and scalable language. * Language is Scala * Understanding the language is difficult. |
| Framework comparison | * Python offers GUI   which is the graphical user interface. | * Integrated with spark core, spark mlib and graphx. * It doesn’t offer any GUI |
| Accuracy | * Naïve Bayes Algorithm – 81%   (similarly can used for any algorithms) | * Naïve Bayes algorithm - 85% |

**8.3.1 Project implementation schedule**

This is concerned with:

* **What activities can produce expected project outputs?**

This means that what are the actual implementation in the project that are to be developed that could actually produce the desired output needed by any company or person, so that the developed project or the system can fulfill the basic requirements.

* **What is the sequence of these activities?**

In what manner or rather what sequence should be followed in obtaining the desired requirements gathered by our research analysis.

* **What is the time frame for these activities?**

Each activity for the completion of the whole system need to be completed within the required time frame.

* **Time control and remedial Action**

Time taken to implement project activities is one measure of successfulness of supervision or monitoring of project implementation. So, the proper care was taken so that the project was developed on time and the report for the capstone was developed and formatted within the given time frame, so that we could prove ourselves to be efficient and smart as developers.

* **Project Activity Time Listing**

It lists the timing of all the activities of a project. It shows how much time is required by each activity/module in android. So, an activity should not take much time which can affect the performance of the whole project.

**8.4 Conversion plan**

This section of the system conversion plan describes the overall conversion strategy. The approach to converting the current system which is manual to the automated system is one that results in a new system database is both comprehensive and easy. A series of automated methods are combined to meet the overall goals of the conversion strategy for a comprehensive and valid system for the organization, these goals are achieved through a cooperative, efficient, and simple approach undertaken by the developers. The success of the conversion strategy is contingent upon assurance that a number of clearly identified prerequisite conditions are met before conversion activities begin at the testing level.

|  |  |  |
| --- | --- | --- |
| **Activities Begin Dates End Dates** | | |
| Planning Meetings | August 10,2018 | August 18,2018 |
| Preparing Plans | August 19,2018 | September 10,2018 |
| Analysis | September 15, 2018 | September 28, 2018 |
| Design | October 1, 2018 | October 15, 2018 |
| Development | October 16, 2018 | October 30, 2018 |
| Unit Testing | November 1, 2018 | November 6, 2018 |
| System Testing | November 7, 2018 | November 12, 2018 |

Table 8.4.0 Conversion Plan

# 9. PROJECT LEGACY

**9.1 Current Status of Project:**

The current status of the TWITTER SENTIMENTAL ANALYSIS**:** The project is fully functional according to the requirements being listed in the SRS during the meeting period. The changes were made into the system after the testing phase so that the system can meet up the needs of the organization or any person to the fullest possible way it can. But still the updating and maintenance will be provided to the user time to time. All modules of the project are developed. Designing is completed and made quirky for the user to get more user experience.

* 1. **Technical and Managerial Lessons Learnt**

**Technical:**

* Learnt something new by working on live twitter data.
* Enhanced our knowledge in machine learning, NLTK tools, python.

**Managerial**

* Learnt Team Management.
* Learnt how to deal with different situations when stuck during the coding and the designing part.
* Learnt how to work in group of developers
* Learnt to distribute tasks to members.
* Learnt about the leadership quality.
* Learnt about the software skills that must be present in an engineer to be a good project manager.

# Summary of work done

After doing regular coding and all the designing and testing from last 4 months now we are ending up with a fully functional modules and GUI

* **Perumalla Murali Mallikarjuna**: This project is divided to 4 different modules. He is responsible for developing code for twitter streaming and classification using machine learning algorithms and pickling concept. He is also responsible for development of GUI for preprocessing and cleaning of data. He has gathered tweets on 3 different mobile products and done the visualization.
* **Mahesh Reddy Baddigam:** In this project we used natural language processing for cleaning the data. He is responsible for developing code cleaning the data. He used offline dataset for hate speech detection and made word clouds on positive and negative tweets. He also took part in machine learning algorithms.
* **Swathy R:** She is responsible for the visualization in IBM Watson (the data was extracted in other phase of the project). Also responsible for the word cloud using R. Responsible for the comparison of Spark with Python to predict which tool give better accuracy. Documented the project and did research on the future scope of sentiment analysis on Twitter where she compared the techniques and frameworks which are currently in market.
* **Swetha R:** She is responsible for the visualization in R Studio (the data was extracted in other phase of the project). As the data extraction is the part of Big Data and R is one of the tools, tried to sentiment analyze with R tools also it helps in better visualization of large amount of data extracted, dividing the data into different Documented the project and did research on the future scope of sentiment analysis on Twitter.
* **Apurva Roshan:** Responsible for the development of GUI interface connecting the visualization and pre-processing part of the project. Done visualization in Python which will import the extracted cleaned tweets from the directory and plotted according to it in the form of GUI interface. Visualization in tableau and the dashboards for comparing different views in tableau. Documentation of the report and did research on the future scope of sentiment analysis on Twitter.

**10. USER MANUAL FOR GUI**

**10.1 Installing Application**

* **Prerequisites**
  + Any operating system with python support
  + File should be csv
  + column name of the tweets should be tweets
  + Minimum configuration of 512 MB RAM and 512 MB of free internal storage.
  1. **. Homepage**
  + Home page for processing and visualization of data
  + Click on preprocessing browse required file for cleaning and analysis
  + Click on clean button to clean the selected files
  + The cleaning file would be stored in present working directory
  + Click on the visualization for the better analysis of extracted cleaned data
  + User can see the data with the help of button Real Time Data
  + User is provided with the option of IBM Watson analysis for online Platform, Python and Tableau for the offline version.

# C:\Users\Swetha\Pictures\Screenshots\Screenshot (196).png11. SCREENSHOTS

Fig 11.0.0 Home Page of GUI

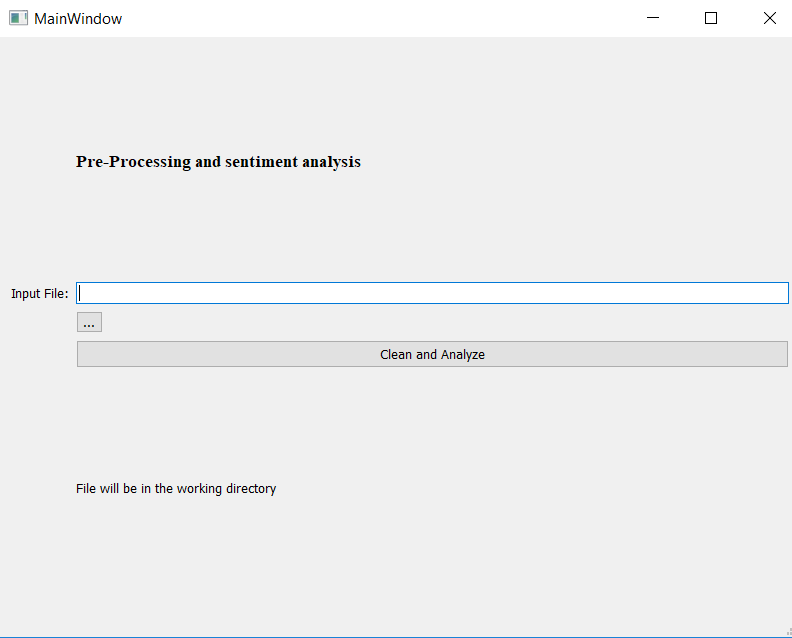


Fig 11.0.1 Pre-Processing stage

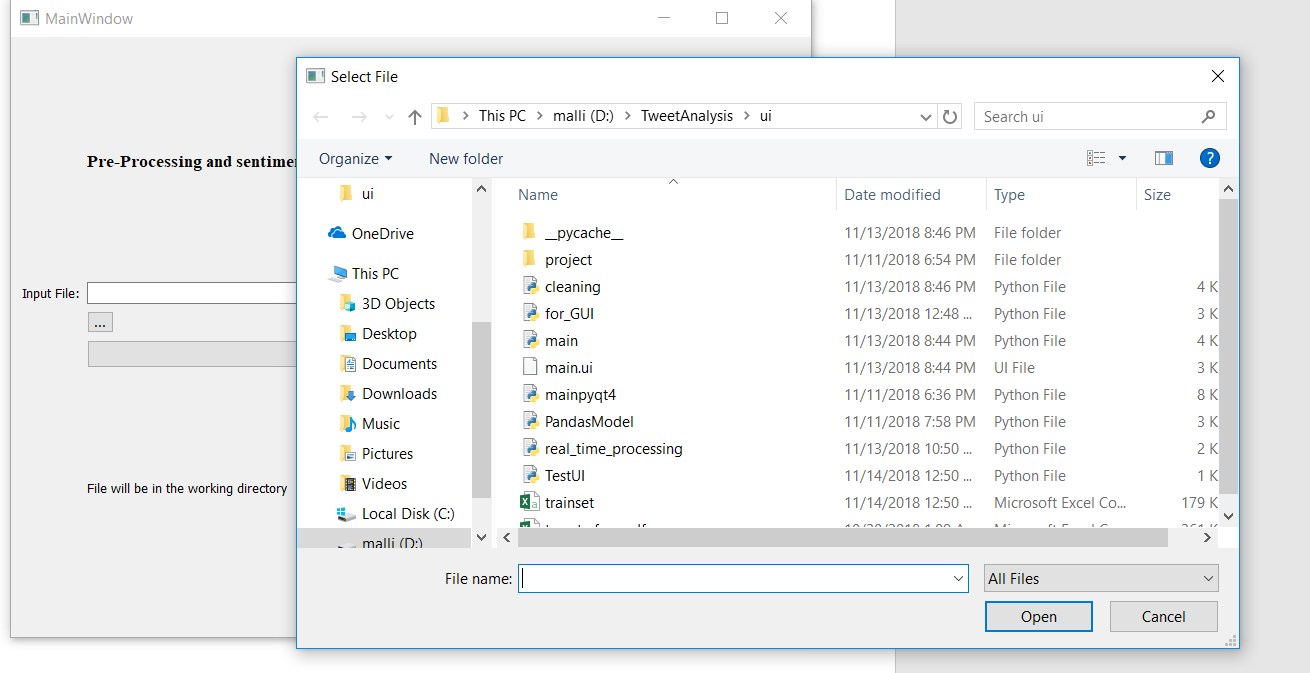
****

Fig 11.0.2User browsing file

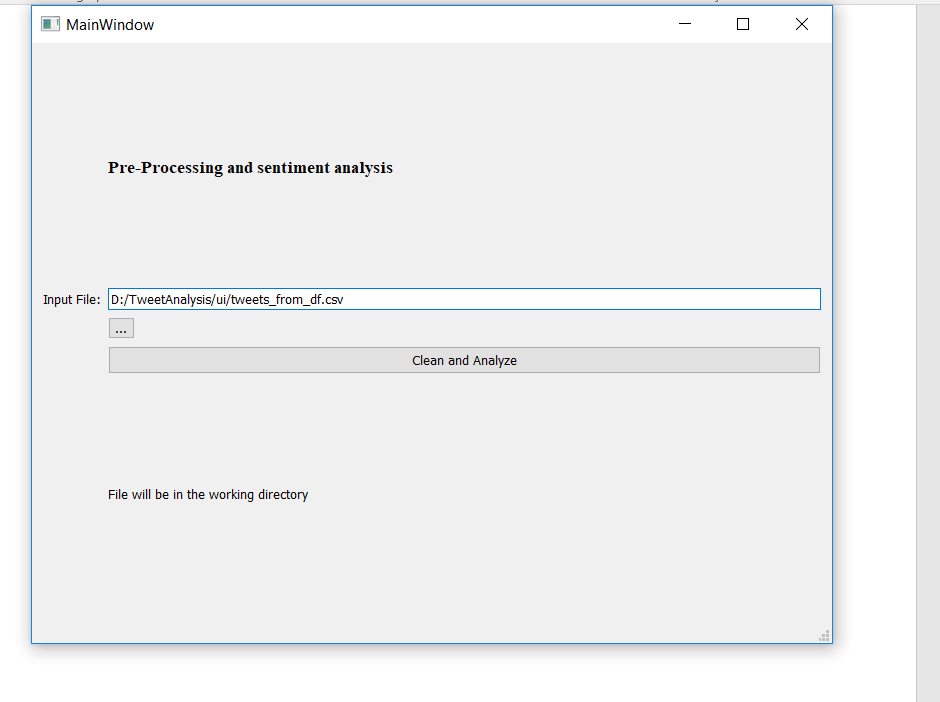
****

Fig 11.0.4 File is selected

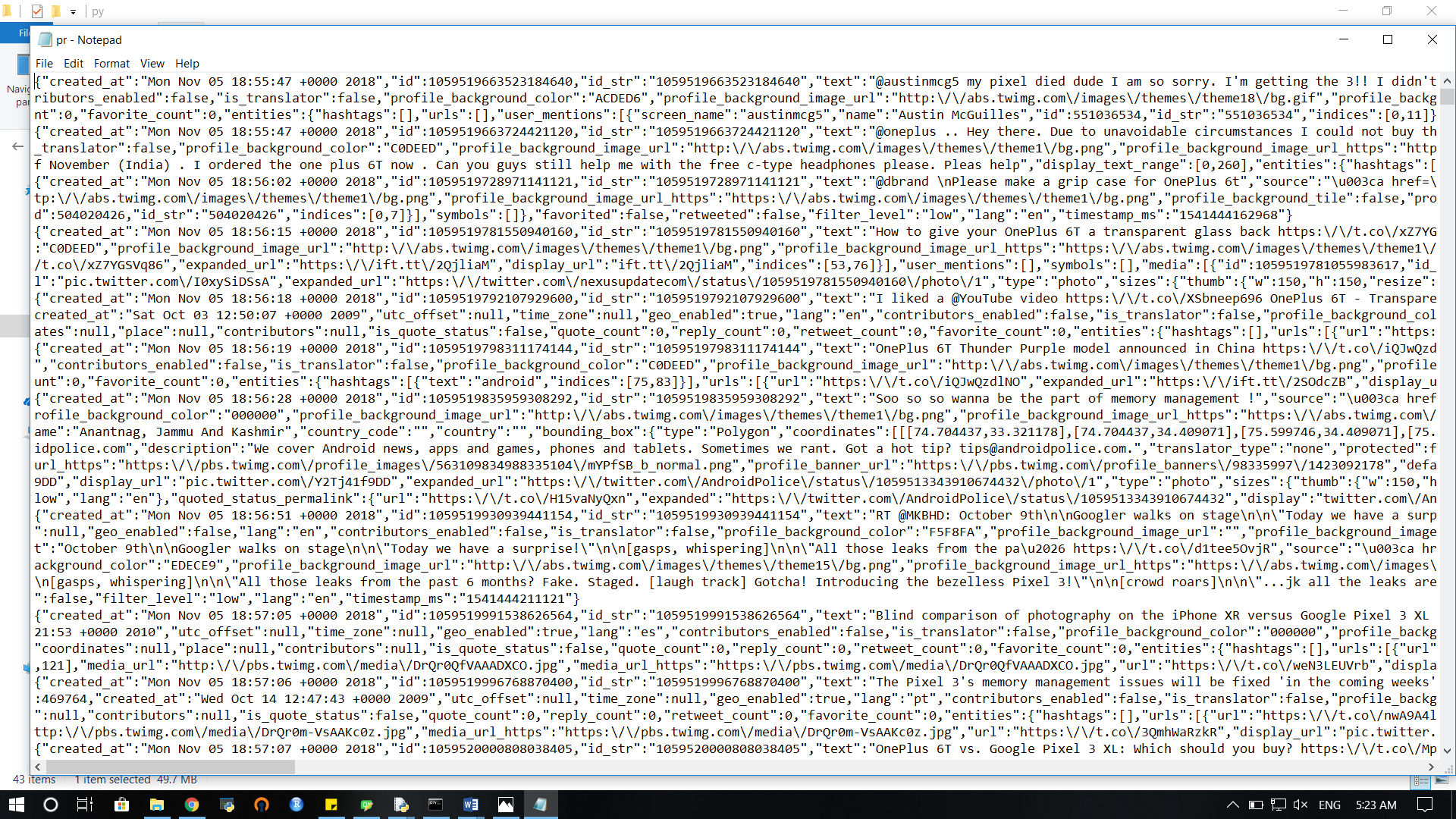
****

Fig 11.0.5 Screenshots of streamed tweet

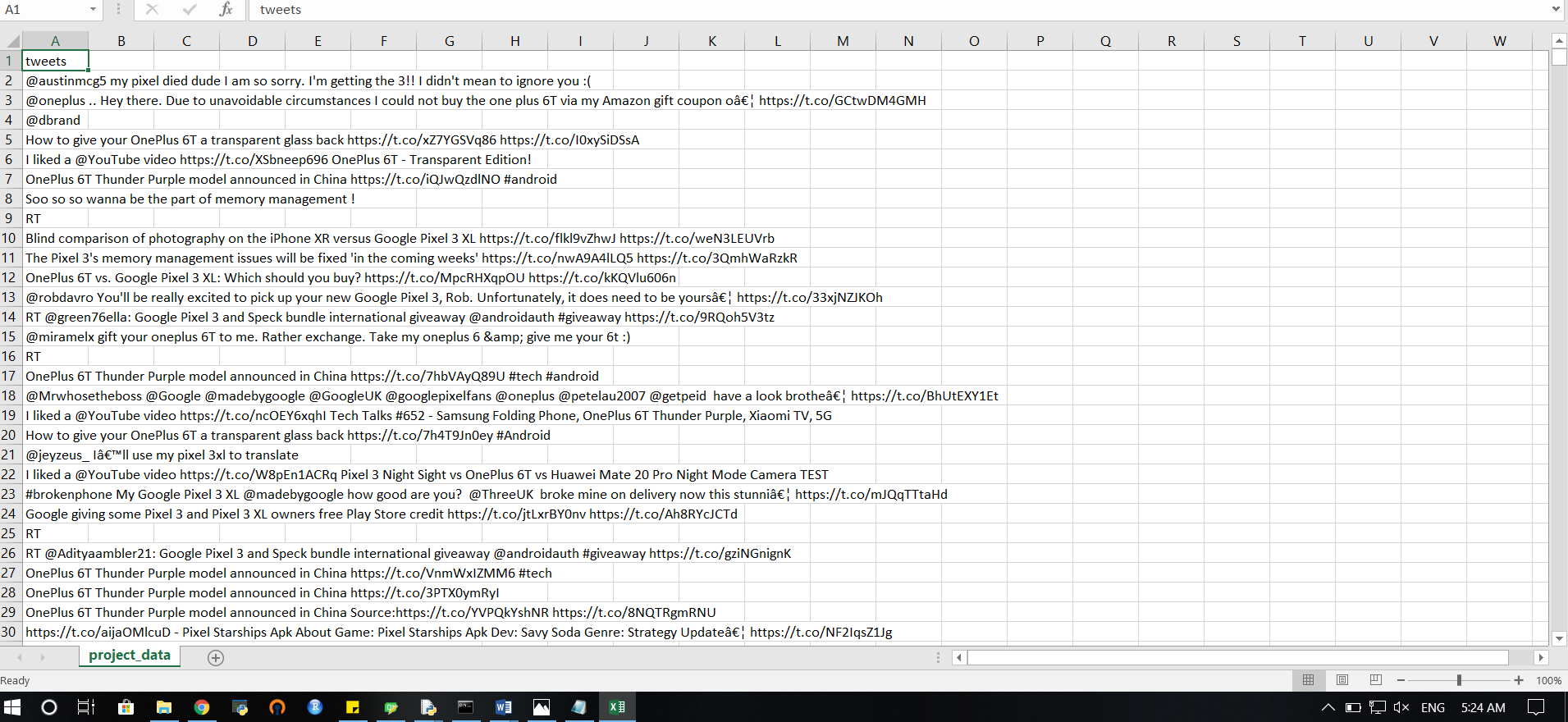
****

Fig 11.0.6 Tweets from json format to csv

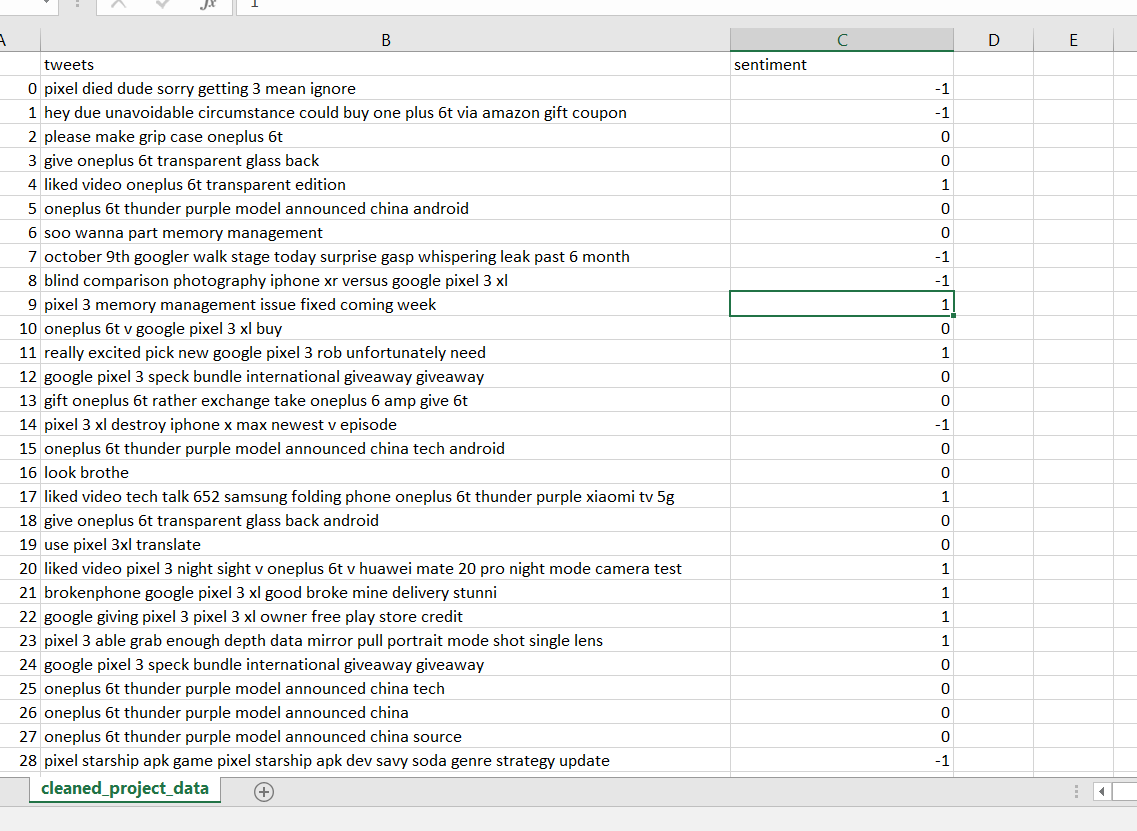


Fig 11.0.7 Cleaned Tweets with sentiment score

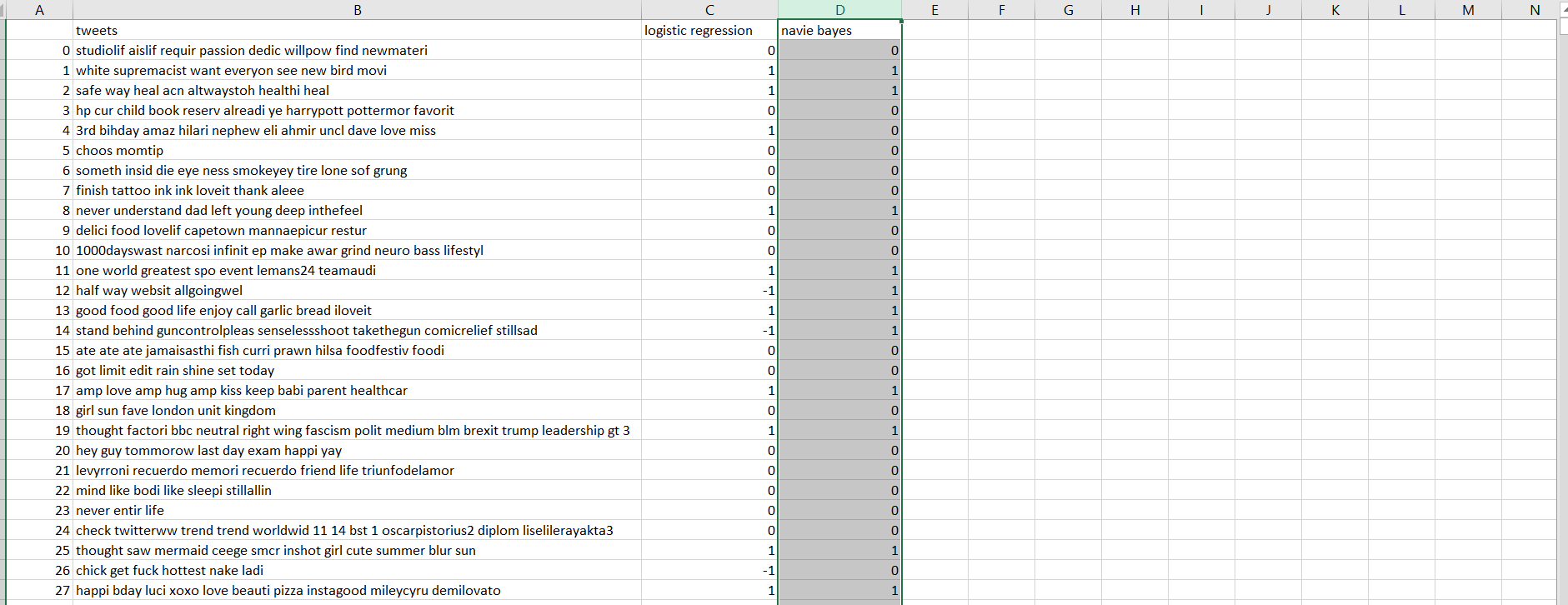


Fig 11.0.8 Screenshot of tweets classified by ML algorithms

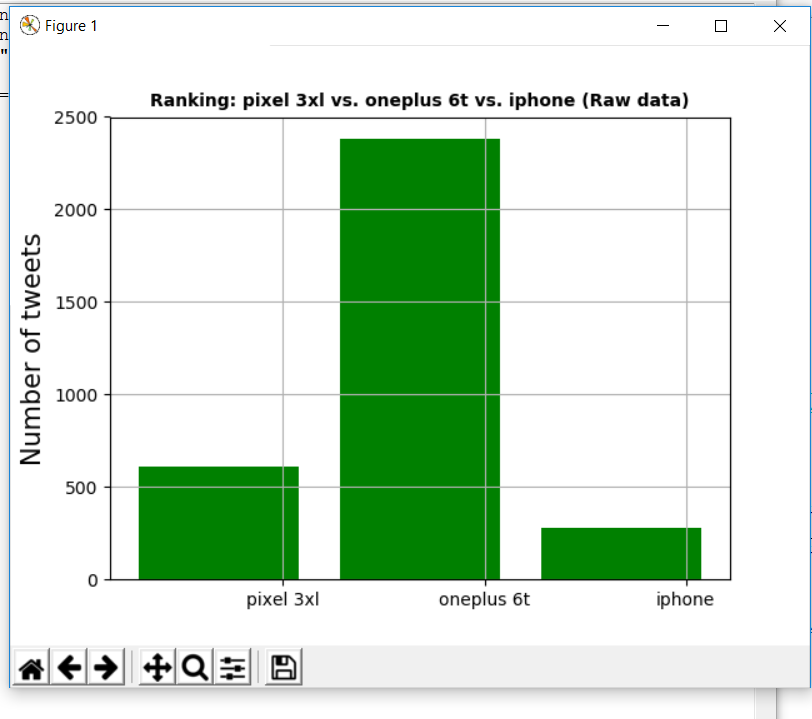
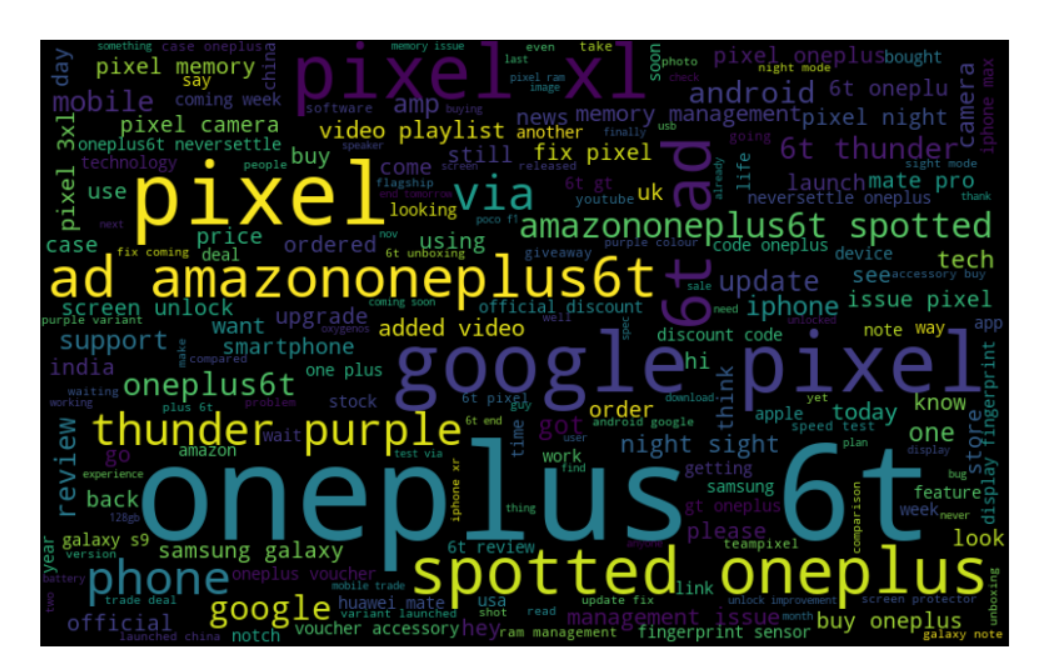


Figure 11.0.9 Screenshots of visualization



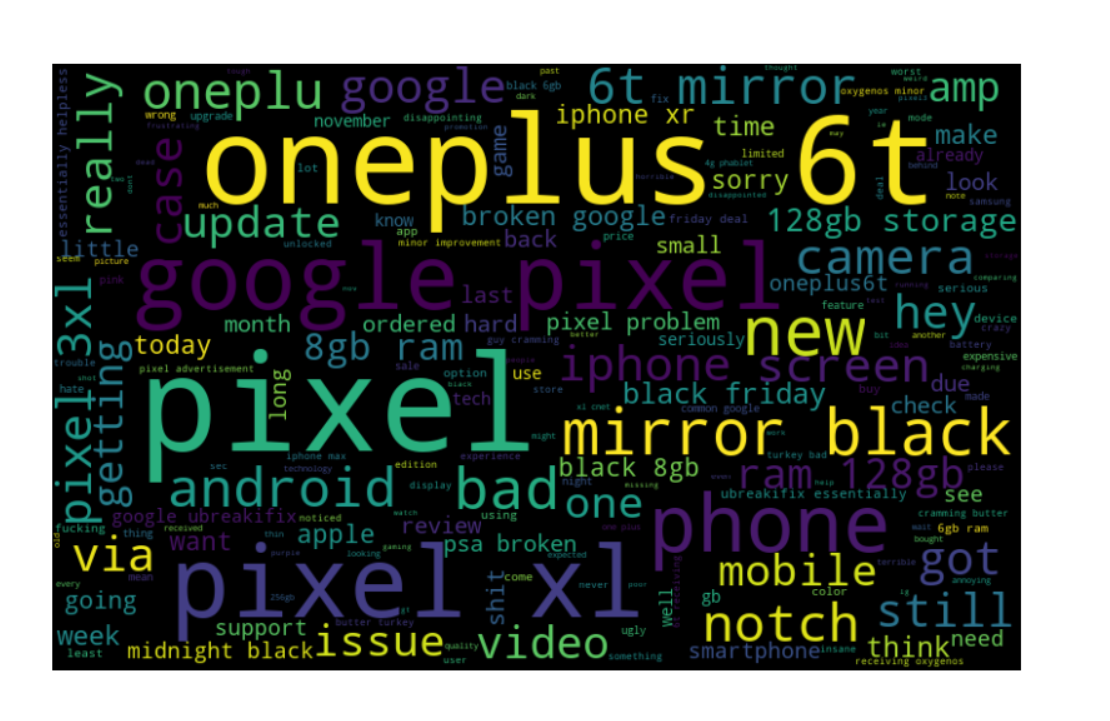
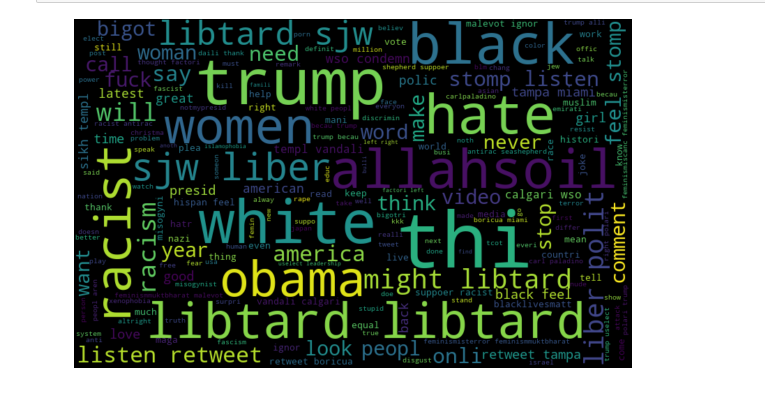


Fig 11.0.10 Wordclouds of positive tweets and negative tweets on mobile dataset



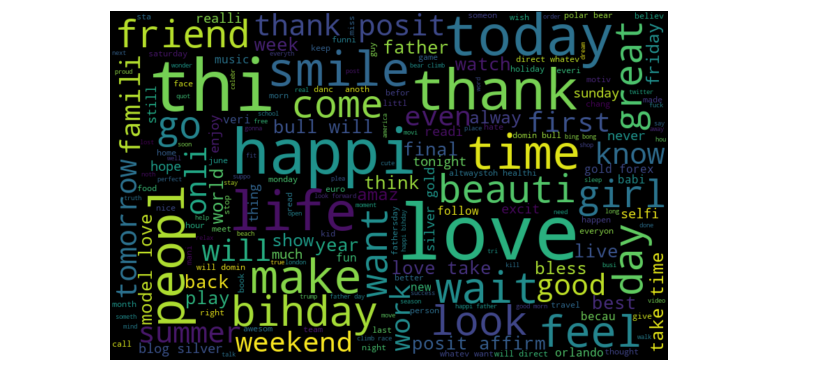
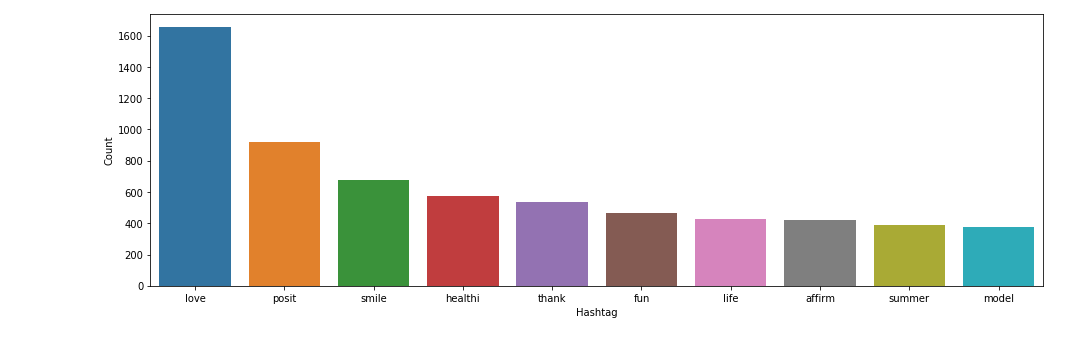


Fig 11.0.11 Wordclouds on hatespeech detection dataset

Fig 11.0.12 Word cloud of positive tweets



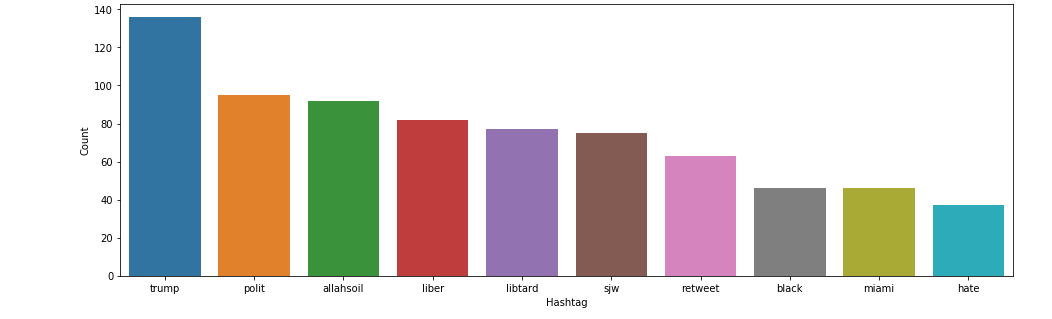


Fig 11.0.13 and Fig 11.0.14

Bar plot of top positive and negative hashtags in hate speech detection

# 

# 12. BIBLIOGRAPHY

# 12.1 Web References

* https://www.sklearn.org.com/
* https://www.pythonprogramming.org/
* https://www.geeksforgeeks.com/
* https://stackoverflow.com
* https://analyticsvidhya.com
* Refer from research paper done by Swathy R, Swetha R, Apurva Roshan