CHAPTER 1

INTRODUCTION

1.1 Introduction to Twitter Data Analytics

Social media like Twitter is a continuous source of real time information. The data present on social media can be used as knowledgeable information to serve various domains like depicting natural calamity (earthquakes etc.), finding the thoughts of people about any person or a particular news/event. Tweet text is a short text message limited by 140 characters in length posted by users on Twitter. People often post messages about interesting news, their daily activities, thoughts, feelings, as well as health conditions.Twitter®1 is a massive social networking site tuned towards fast communication. More than 140 million active users publish over 400 million 140-character "Tweets" every day2. Twitter's speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events, such as the Arab Spring3 and the Occupy Wall Street movement4. Twitter has also been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy.

Twitter is a "micro-blogging" social networking website that has a large and rapidly growing user base. Thus, the website provides a rich bank of data in the form of "tweets," which are short status updates and musings from Twitter's users that must be written in 140 characters or less to tell others what they are doing, what they are thinking, or what is happening around them. According to the latest Twitter entry in Wikipedia, the number of Twitter users has climbed to 190 million and the number of tweets published on Twitter every day is over 65 million .As an increasingly popular platform for conveying opinions and thoughts, it seems natural to mine Twitter for potentially interesting trends regarding prominent topics in the news or popular culture.

Tweets and status updates range from important events to inane comments. Most messages contain little informational value but the aggregation of millions of messages can generate important knowledge. Several Twitter studies have demonstrated that aggregating millions of messages can provide valuable insights into a population. Barbosa and Feng (2010) classified tweets by sentiment, a first step towards measuring public opinion, such as political sentiment, which has been shown to track public political opinion and predict election results (Tumasjan et al. 2010; O'Connor et al. 2010). Eisenstein et al. (2010) studied lexical variations across geographic areas directly from tweets. Others have monitored the spread of news (Lerman and Ghosh 2010), detected the first mention of news events (Petrovic, Osborne, and Lavrenko 2010), and moni- ´ tored earthquakes (Sakaki, Okazaki, and Matsuo 2010). Copyright c 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. Twitter users often publicly express personal information; messages like "I got da flu" and "sick with this flu it's taking over my body ughhhh" are common. Knowing that a specific user has the flu may not be interesting, but millions of such messages can be revealing, such as tracking the influenza rate in the United Kingdom and United States (Lampos and Cristianini 2010; Culotta 2010b). Furthermore, tweets are not isolated events: they occur with specific times, locations, languages and users. Aggregating over millions of users could provide new tools for research.

1.2 Rapid Miner

Rapid Miner is the most powerful, easy to use and intuitive graphical user interface for the design of analytic processes. Hundreds of data loading, data transformation, data modelling, and data visualization methods with access to a comprehensive list of data sources including Excel, Access, Oracle, IBM DB2, Microsoft SQL, Netezza, Teradata, MySQL, Postgres, SPSS, Salesforce.com, and hundreds more! Easily integrate your own specialized algorithms into Rapid Miner by leveraging its powerful and open extension APIs. Rapid Miner Studio runs on every major platform and operating system. With more and more employees bringing their own devices, you need flexibility in how and where your analytics are processed. Rapid Miner Studio breaks away from the limitations of traditional data analysis tools and allows you to work with large data sources.RapidMiner (formerly YALE)
is the world-wide leading open-source data mining solution due to the combination of its l eading-edge technologies and its functional range.
Applications of RapidMiner cover a wide range of realworld data mining tasks. The mod ular operator concept of RapidMiner (formerly YALE) allows the design of complex nest ed operator chains for a huge number of learning problems in a very fast and efficient wa y (rapid prototyping). The data handling is transparent to the operators. They do not have to cope with the actual data format or different data views - the RapidMiner core takes car e of all necessary transformations.

1.3 Twitter 4J

Twitter4J is an unofficial Java library for the Twitter API.With Twitter4J, you can easily integrate your Java application with the Twitter service. Twitter4J is an unofficial library.

Twitter4J is featuring:

✓ 100% Pure Java - works on any Java Platform version 5 or later
✓ Android platform and Google App Engine ready
✓ Zero dependency : No additional jars required
✓ Built-in OAuth support
✓ Out-of-the-box gzip support
✓ 100% Twitter API 1.1 compatible

1.4 JxBrowser

JxBrowser is a cross-platform Java library that allows integrating Google Chromium-based web browser component into the Java Swing/AWT/JavaFX application. With JxBrowser we can embed a lightweight Swing/JavaFX component into our Java application to display modern web pages, supporting the latest web standards such as HTML5, CSS3, and JavaScript etc. JxBrowserPanel is a component that currently has the sole purpose of "host" for the JxBrowser. It is a class that inherits from JPanel and only has the BrowserView the JxBrowser inside. To use this component we need only jxbrowser.

JXBROWSER FEATURES:

- **Rich and easy-to-use** API.
- **Great** documentation with many examples.
- **You're in control of your app.** Compared to other similar products, JxBrowser does not execute native code in your Java application process. Forget about memory usage issues caused by native web browser engine that allocates too much memory of your Java process. All the native code runs in separate native processes. In the unlikely event of a browser crash (e.g. because of a plugin problem), your app will continue working and you can restore the browser back.

Great support. Responsive support team will answer your request within 24 hours.

1.5 Google API

Google APIs is a set of APIs developed by Google which allow communication with Google Services and their integration to other services. Examples of these include Search, Gmail, Translate or Google Maps. Third-party apps can use these APIs to take advantage of or extend the functionality of the existing services.

The APIs provide functionality like analytics, machine learning as a service (the Prediction API) or access to user data (when permission to read the data is given). Another important example is an embedded Google map on a website, which can be achieved using the Static maps API, Places API or Google Earth API.

Common uses

- **User registration** is commonly done via Google+ sign in, which allows users to securely log in to 3rd party services with their Google+ account using the Google+ API. This is currently available from within Android, iOS or JavaScript.It is popular to include a "Sign in with Google" button in Android apps, as typing login credentials manually is time-consuming due to limited screen size. As the user is usually signed into their Google account on their mobile device, signing-in/signing-up for a new service with a Google is usually a matter of a few button clicks.

- **Drive apps** are various web applications (often third party) which work within Google Drive using the Drive API. Users can integrate these apps into their Drive from the Chrome Web Store which allows them to do work entirely in the cloud.[9] There are many apps available for collaborative document editing (Google Docs, Sheets), picture/video editing, work management or for sketching diagrams and workflows.

- **Custom Search** allows web developers to provide a search of their own website by embedding a custom search box and using the Custom Search API. They can customize the search results and make money off the ads shown using AdSense for Search.

- **App Engine apps** are web apps that run on the Google App Engine, a platform-as-a-service (PaaS) cloud computing platform which allows web developers to run their

websites in Google datacentres. These web apps often take advantage of APIs to manipulate services such as TaskQueue (a distributed queue), BigQuery (a scalable database based on Dremel) or DataStore.

- **Gadgets** are mini-applications built in HTML, JavaScript, Flash and Silverlight that can be embedded in webpages and other apps. They can run on multiple sites and products (even writing them once allow users to run them in multiple places).

1.6 Frequency Visualization

Cirrus is a word cloud displaying the frequency of words appearing in a corpus. Words occurring more frequently appear larger. Cirrus is a freely available visualization tool that generates a word cloud from a document or group of documents. The initial word cloud contains all of the most frequent words from the text, laid out graphically to fit together within an elliptical shape. Words may be oriented either vertically or horizontally, and they are rendered in different colours for ease of viewing. The user should be aware that word orientation and colour are strictly decorative elements, they do not indicate anything meaningful about the word. The size of the word, however, indicates the frequency with which it appears in the document. The larger the font size, the more frequent the word. The user can mouse over a word to see the precise number of times that it appears in the document. One of the problems with word frequency visualization is the prevalence of conjunctions and articles in most texts. These less semantically interesting words often drown out the adjectives, adverbs, nouns, and verbs that are more likely to be of interest to the researcher. Cirrus offers an easy solution to this problem. By clicking the options icon in the upper right hand menu, a user can apply a list of stop words to the visualization. These stop words will be treated as noise, and stripped from the cloud, allowing words of greater interest to surface. Cirrus offers two pre-built lists of stop words, one for common English words, and one for common French words. The user can view the lists, so she know which words will be removed before applying them.

1.7 Maptive

Maptive is an online tool Transform raw location data into a beautiful, customized Google map in a matter of seconds.

FEATURES:

- Large or small data sets, 10 or 100,000 locations.
- Safe & Secure
- Cloud-Based
- Get access to all the customization and editing tools you'll ever need to create beautiful, customized maps.
- Create territories that make sense and improve the profitability of your business.
- Precision level control makes it easy to filter data and show only what you want to see on the map you create and customize.
- Optimize Routes / Directions
- Fully customized and configured to meet the unique needs of any business or organization—large or small.
- Share your map privately with specific individuals, publish publicly to the web, or embed within a webpage or blog using simple HTML code.
- The maps you create and share can be easily accessed and viewed on any smartphone or tablet devices.
- Track Your Location with GPS

1.8 Plotly

Plotly, also known by its URL, **Plot.ly**, [1] is an online analytics and data visualization tool, headquartered in Montreal, Quebec. Plotly provides online graphing, analytics, and stats tools for individuals and collaboration, as well as scientific graphing libraries for Python, MATLAB, Perl Julia, Arduino, and REST

Plotly has a graphical user interface for importing and analysing data into a grid and using stats tools. Graphs can be embedded or downloaded. Mainly used to make creating graphs faster and more efficient.

Plotly is the easiest way to graph and share your data. It standardizes the graphing interface across scientific computing languages (Python, R, MATLAB, etc.) while giving rich interactivity and web share ability that has not been possible before with matplotlib, ggplot, MATLAB, etc.
On the Plotly website, you can style your graphs with a GUI, so you don't have to spend hours writing code that simply changes the legend opacity.
Plotly does this all while backing up your graphs on the cloud, so that years later, you can find data that may have otherwise been on a harddrive in a landfill. If you make your data public, other people can also find your graphs and data. The best practice that we have today for saving and sharing research data is to entomb it as a thesis in the engineering library basement. All that is changing.

CHAPTER 2

TWITTER DATA ANALYTICS

2.1 INTRODUCTION TO TWITTER DATA ANALYTICS

Social media sites like Twitter, Facebook helps to communicate and share information in mass .As a result, large and tons of data is present which can be used to discover information related to any news/event, person or any health issue. This information can thus be used to serve many purposes like:

- Developing surveillance system
- Mining information etc.

Twitter Data Analytics is divided into three categories:

- News Analysis
- Celebrity Analysis
- Health Analysis

2.1.1 INTRODUCTION TO NEWS ANALYSIS

The news media, and more specifically print media, serve as valuable sources of information and powerful modes of communication. This power controls much of what people understand of events that occur around the world on a daily basis. The way information is transferred to its recipients comes through various forms of communication, all of which is framed to meet the goals of the providing source. The aim of this project is to find out the thoughts of people about any news or event occurred. This is done using Rapid Miner, a java based software for extracting the tweets, processing it and analysing the feeling of the people by categorizing that into three categories: positive, negative and neutral

2.1.2 INTRODUCTION TO CELEBRITY ANALYSIS

Social media sites like Twitter, Facebook helps to communicate and share information in Mass .As a result, large and tons of data is present which can be used to discover Biomedical and health related information. This information can thus be used to serve Many purposes like:

- To analyse behaviour of any person about him/her
- To analyse behaviour of other persons towards any him/her
- To analyse follower's tweets.
- Geomapping of follower's geolocation
- Analyses of two twitter data analysing tools which are Tweetchup and Twitnomy.

Our aim to analyse the data present on social media. People often post messages about Interesting news, their daily activities, thoughts and feelings. This may help to analyse Their behaviour and behaviour of other people's towards him/her. It makes easy to analyse Positive and negative attitude from their tweets. Celebrity's publicity or how

famous they are, what people think about them and how positive and negative image they have.
This can be analysing from tweets. Using their follower's tweets and geolocation it can be Easily analyse their behaviour and how their popularity varying country wise or any Specific area wise. It will also help to find terror area connection by getting number of Followers lying in that area. Twitter message, when used in combination, helps increase The accuracy of our prediction. The data is also verified by analysing twitter data using Twitter data analysing tools which are twitonomy**&**tweetchup.

The basic need for social networking behaviour analyzing is to analyse the data posted into
Twitter and to transform those tweet related data into related information for tweet Analysis. Also to fetch follower's information and any particular person's information and
Relate it to the required information in order to make relations among these info and to Mine the behaviour related data so as to analyze positive and negative attitude from their Tweets. Celebrities publicity or how famous they are, what people think about them and How positive and negative image they have, all this can be analyze from tweets. Twitter Message, when used in combination, helps increase the accuracy of our prediction.

2.1.3 INTRODUCTION TO HEALTH ANALYSIS

Twitter is a great data resource to monitor these health issues prevailing among general population. This project includes the extraction of tweets using Twitter API and then analysing them in many possible ways. Geo mapping is one of the important features explored in this project. Static Geo-Mapping and Dynamic Geo-Mapping are the two types which have been discussed. Apart from this, a tool named Cirrus has been used to take the analysing process to another step. Our aim is to handle the tweets on Twitter relating to health related issues and then extract the required information for analysis.

2.2 NEED FOR TWITTER DATA ANALYTICS

The basic need for twitter data analytics is to analyse the data posted into Twitter and to transform those tweet related data into related information for tweet analysis. Also to fetch follower's information and relate it to the required information in order to make relations among these info and to mine the related data.

2.3 OBJECTIVE OF TWITTER DATA ANALYTICS

Our aim is to handle the tweets on Twitter relating to any news, person or health related issues and then extract the required information for analysis. This project includes the extraction of tweets using Twitter API and Rapid Miner and then analysing them in many possible ways. Geo mapping is one of the important features explored in this project. Static Geo-Mapping and Dynamic Geo-Mapping are the two types which have been discussed. Apart from this, a tool named Cirrus has been used to take the analysing process to another step.

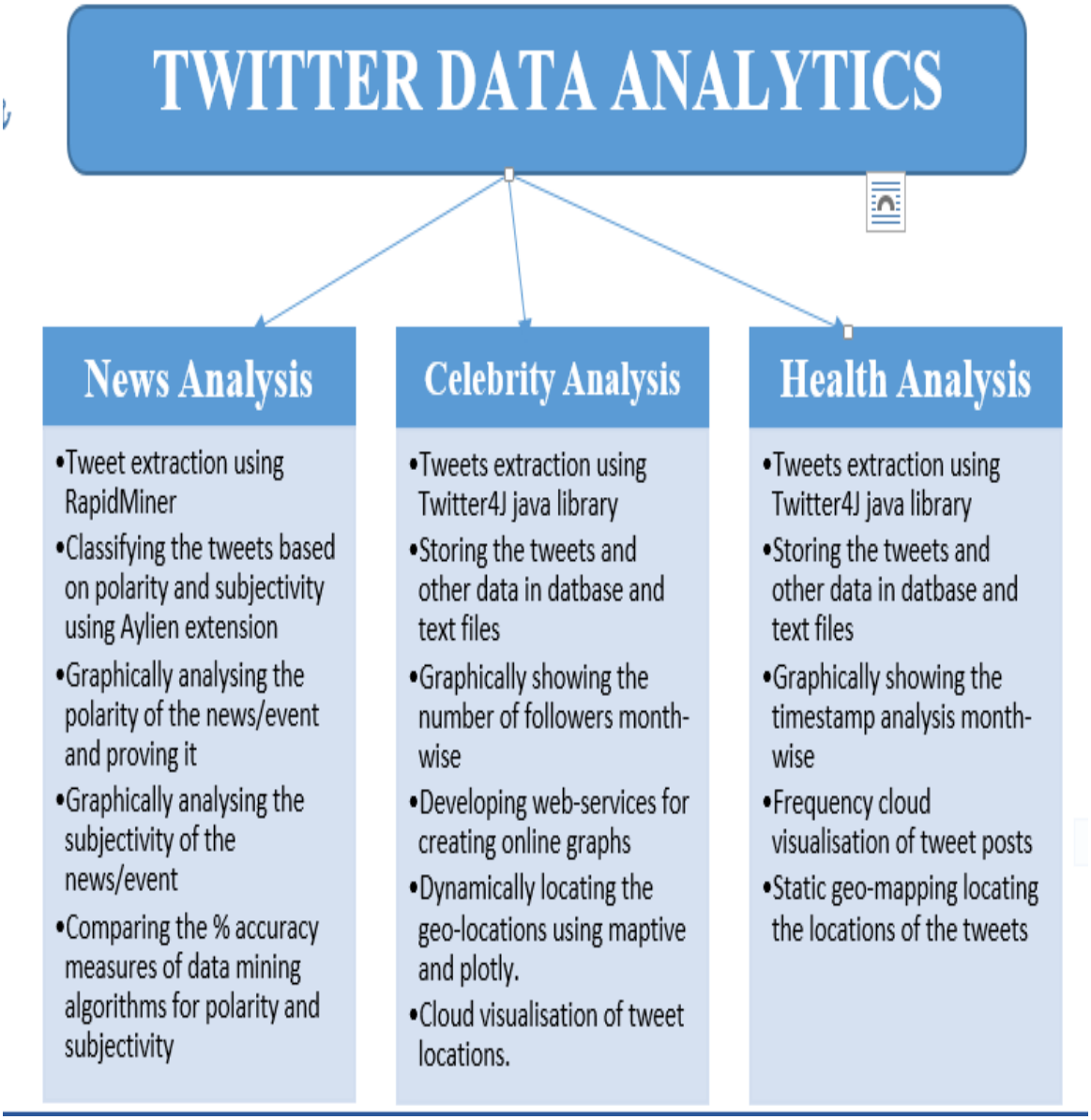

Figure 1: Flowchart showing the Twitter Data Analytics

CHAPTER 3

DESIGNING AND IMPLEMENTATION OF TWITTER DATA ANALYTICS

In this project we present a more general approach that discovers associations from tweets. This project first aims at extracting tweets using two different approaches. Various tweets' attributes like username, language of the tweet, month, year and location from where it was posted are extracted along with the tweet messages. These attributes, along with the original tweet message are then stored in various formats. Using this data from database, location frequency is calculated for every country, which is then used by JxBrowser and Google API service of maps marker for Geo-Mapping. Cirrus, another tool used for analysing the extracted data, is a word cloud displaying the frequency of words appearing in a corpus. This tool not only allows us to remove the stop words, but it also helps us to analyse the tweets by calculating the frequency of important and relevant words.

3.1 Tweets Extraction and Collection

This phase is further categorised as per the three categories of our project.

3.1.1 Tweets Extraction and Collection in News Analysis

Using Search Twitter operator in Rapid Miner, tweets can be extracted and collected in table form. The table consists of columns as: from user, to user, text, relocations, source, language, rewet count.

Figure 2 shows an example on how to gather tweets in Rapid Miner. Below we're searching for tweets containing the keyword "Samsung" using Search Twitter Operator. We've cleaned up our search a little by removing rewets (-rat) and links (-http). We've also restricted the number of tweets to collect to 20 and decided we only want to see English tweets by adding "en" in the language parameter. We've also indicated that we want only recent or popular tweets to be returned using the Result type parameter.

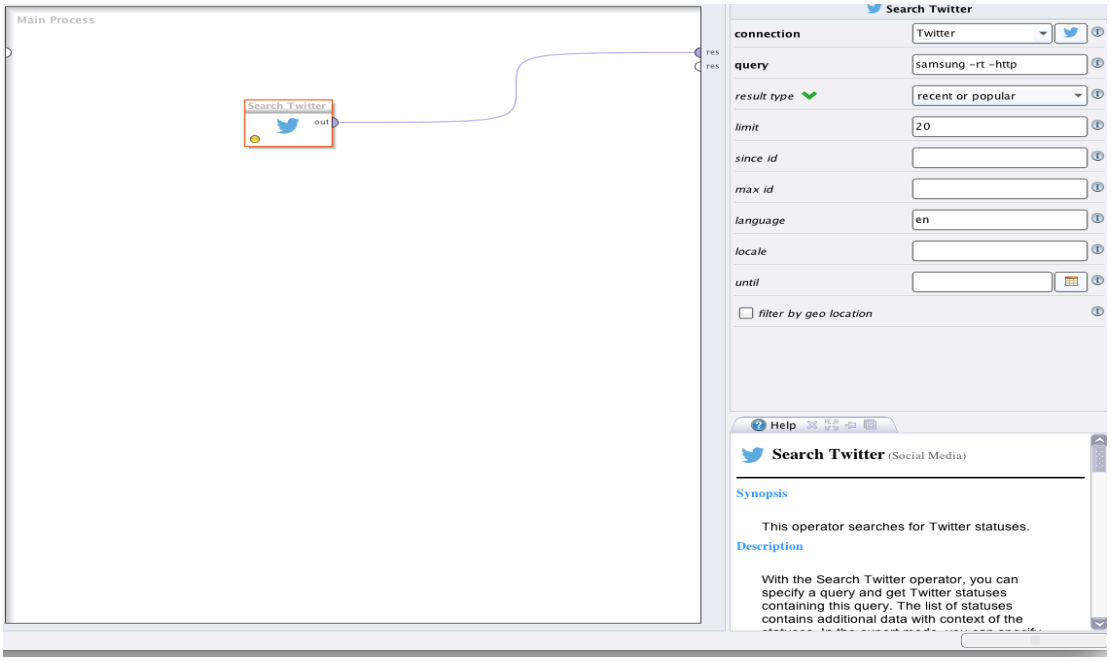

Figure2: Extraction of tweets

After running the process as shown in figure1, the output screen is shown in figure 3 which is in table form consisting of multiple columns consisting of username,user-id,text,source,retweet count etc. proving various kinds of data to be analysed further.

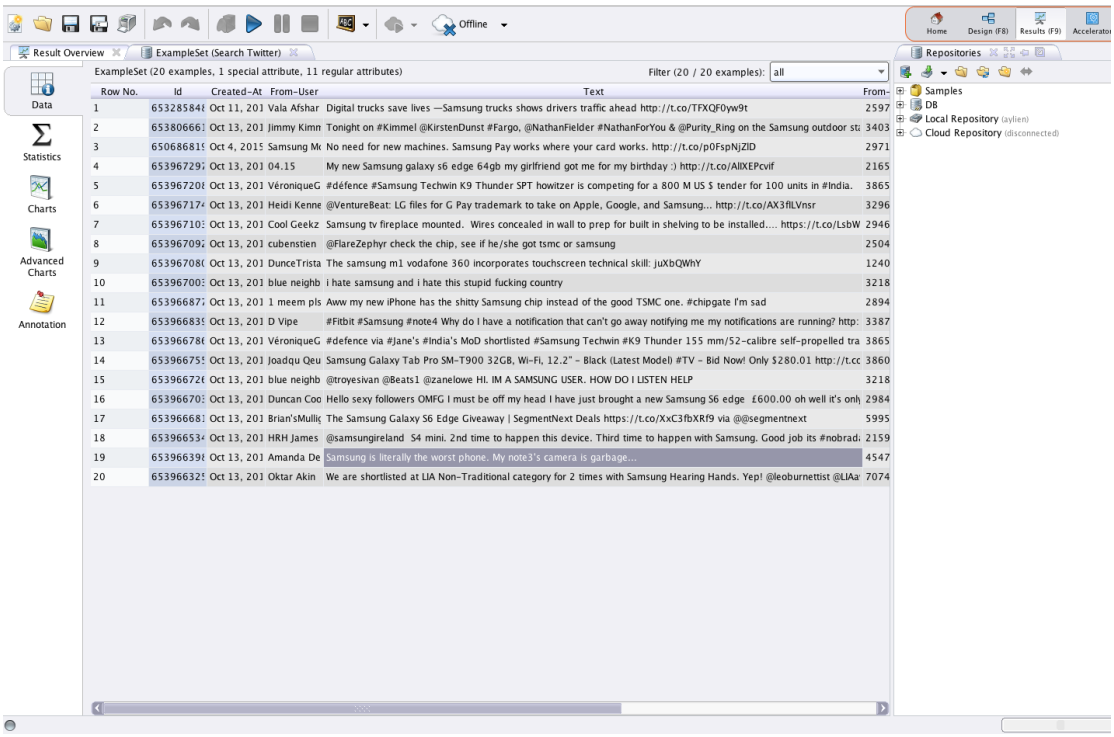

Figure 3: Example Set of gathered tweets

3.1.2 Tweets Extraction and Collection in Celebrity Analysis

The developer registers on twitter to access APIs.the twitter then issues customer token and secret then the application directs user to twitter to verify user credentials where the user enters the required credentials, these credentials are verified by twitter and if correct then twitter issues a OAuth verifier. After this the developer requests access token through the application and twitter issues him the access token and the secret key so that the user can get access to the requested contents and information.

Figure 4 represents that twitter has issued a authentication url to verify credentials of user and figure 5 shows that user has been verified by authorizing app.

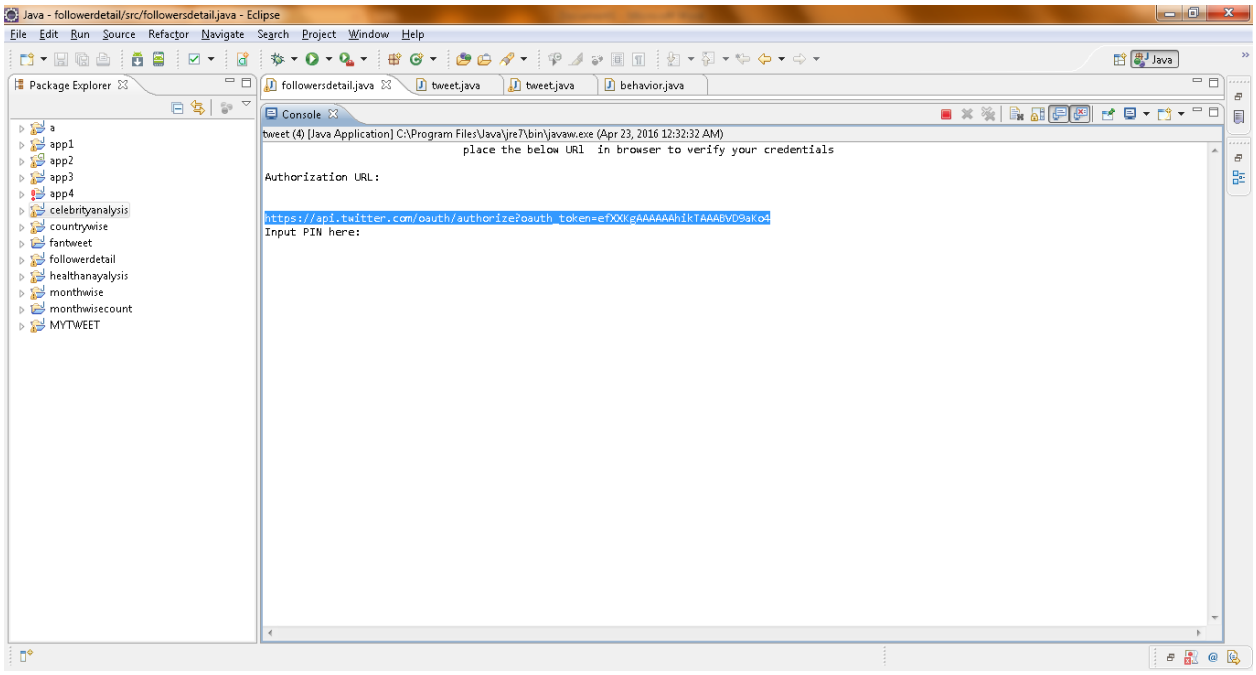

Figure 4: Authentication URL

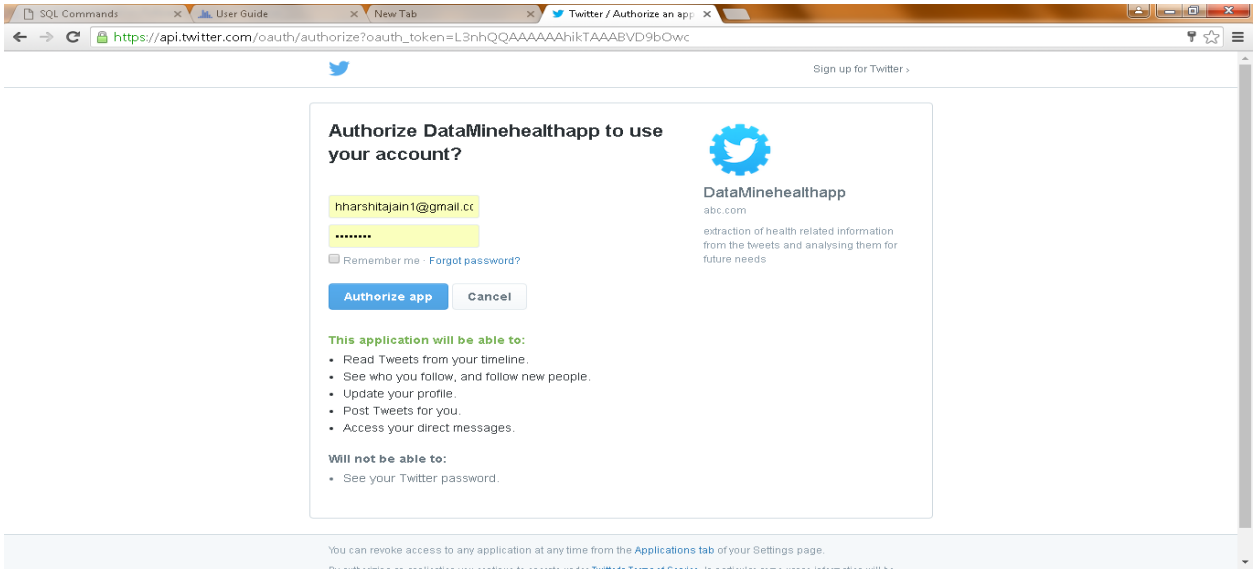

Figure 5:user credential are verified

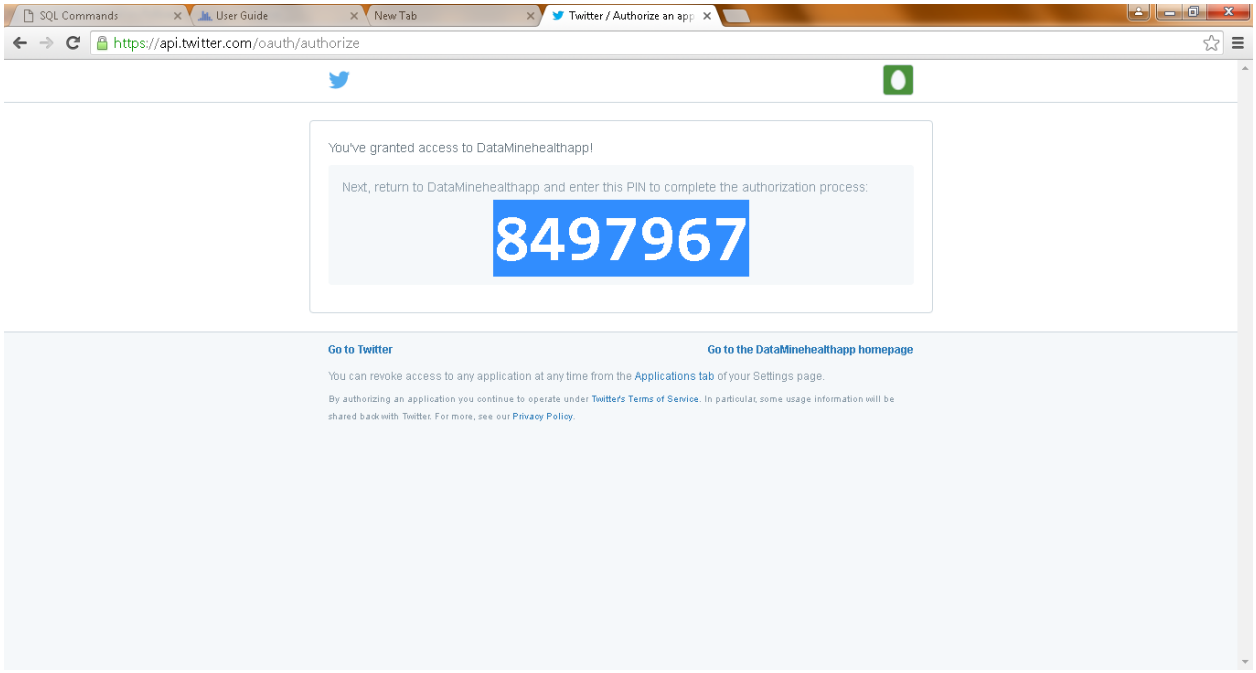

Figure 6: Pin for authentication

The user/developer can get access to his application through the customer token and secret keys. Using these keys he access twitter to grant him the authentication pin which here is 8497967(figure 6) so as to verify the application on twitter .once the user verifies the application he/she has to enter the keyword on the console..

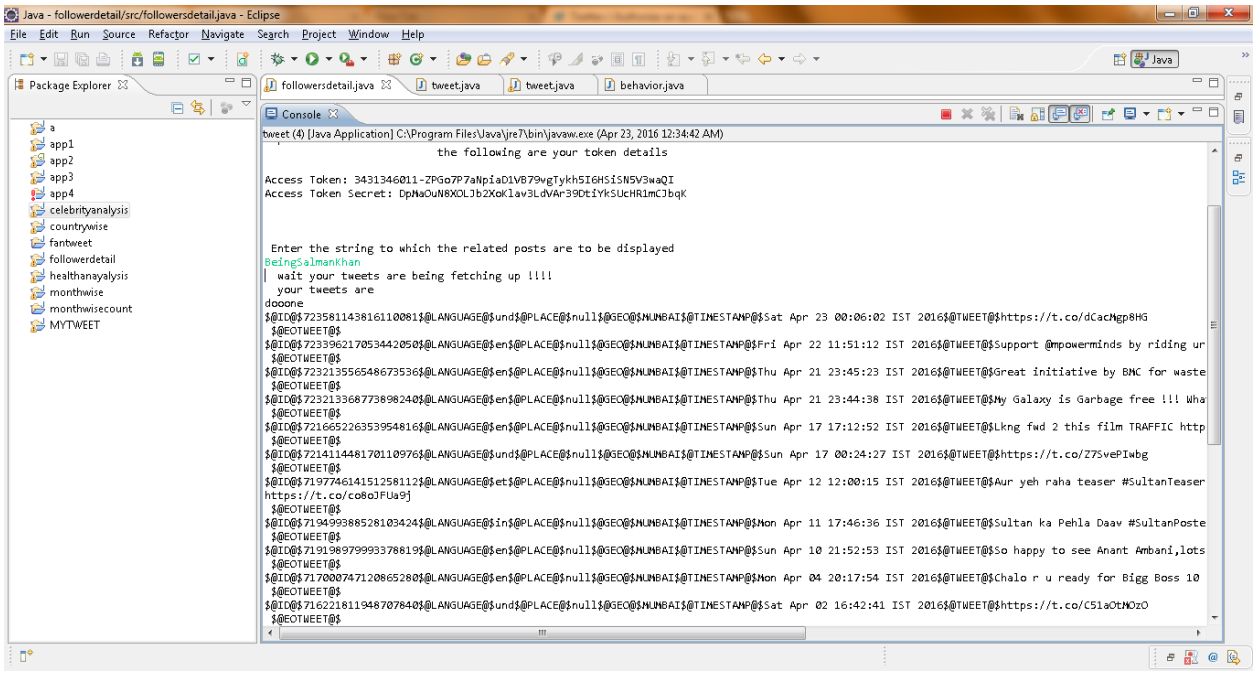

Figure 7: keyword"BeingSalmanKhan" entered

After the authorization and authentication process the user/developer has to specify the keyword or twitter user id in the console output (as shown in figure 7) so that only those tweets and its related information can be collected in which the keyword or from that user id is present. Then once the specification of keyword is done then tweets get extracted on

the console itself and other information is stored in database.

3.1.3 Tweets Extraction and Collection in Health Analysis

A tweet has tweet text, user name, time-stamp, and location. The data pre-processor module reads the raw tweet data from the database, extracts the tweet text, tweet timestamp and user location, and stores the data back into the database.

A language constraint was applied while extracting tweets in order to get tweets of only one particular language, well known to us for better analysis of tweets. Once this constraint was successfully applied, the information attributes such as the username, location, timestamp and the tweet message posted were stored in a database. This was done as a step to make a move further for our analysis and research work.

The data collector module continuously downloads the tweet data using Twitter Streaming API. Twitter Streaming API allows high-throughput near real-time access to global stream of public tweets that matches pre-specified filter predicates. Multiple parameters may be specified in a single connection to the streaming API to determine what tweets will be delivered on the stream. Firstly the tweets were extracted using twitter4j java library. Twitter4J is an unofficial Java library for the Twitter API. With Twitter4J, we easily integrated our Java application with the Twitter service.

Figure 8 shows the steps for extraction of tweet related data and collecting the information in database. Firstly the tweets were filtered on the basis on language constraint in order to extract tweets where language of the follower is English using “en” constraint where “en” stands for English language. After the application of language constraints, the tweet related information such as username, his id number, language, timestamp, geolocation and his/her posted tweet were extracted. Then the tweets were saved in text file and other information that i.e. username, his id number, language, timestamp, geolocation were stored in database so that sql queries can be applied for analysis.

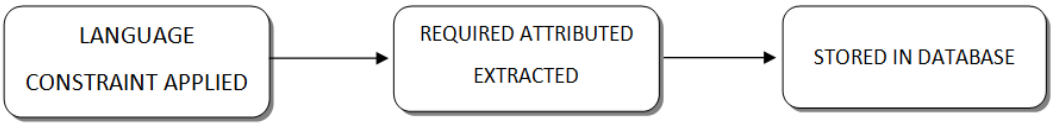

Figure 8: Steps for tweet extraction and collection

Figure 9 shows relationship amongst the twitter application, twitter and the user. With the help of application, the developer registers on twitter to access APIs.the twitter then issues customer token and secret then the application directs user to twitter to verify user credentials where the user enters the required credentials, these credentials are verified by twitter and if correct then twitter issues a OAuth verifier. After this the developer requests access token through the application and twitter issues him the access token and the secret key so that the user can get access to the requested contents and information.

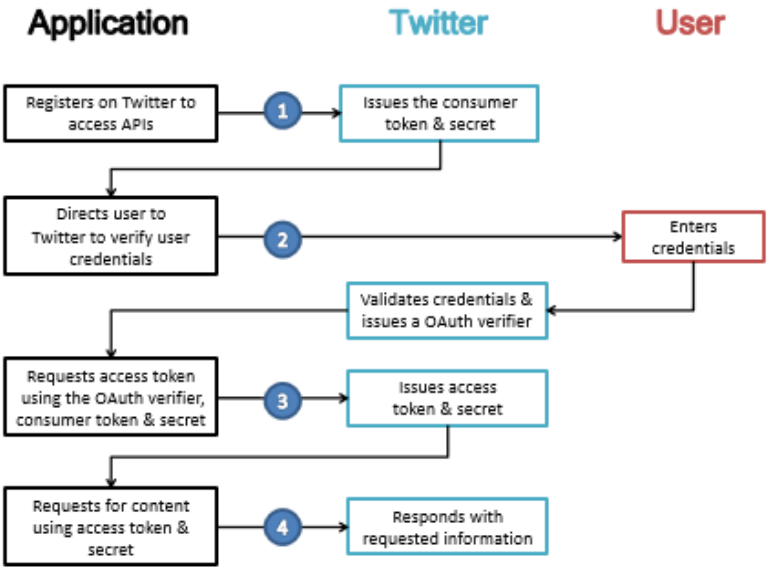

Figure 9: Integration of Java Application with Twitter Services

The user/developer can get access to his application through the customer token and secret keys. Using these keys he access twitter to grant him the authentication pin which here is 1452934(figure 11) so as to verify the application on twitter .once the user verifies the application he/she has to enter the keyword on the console..

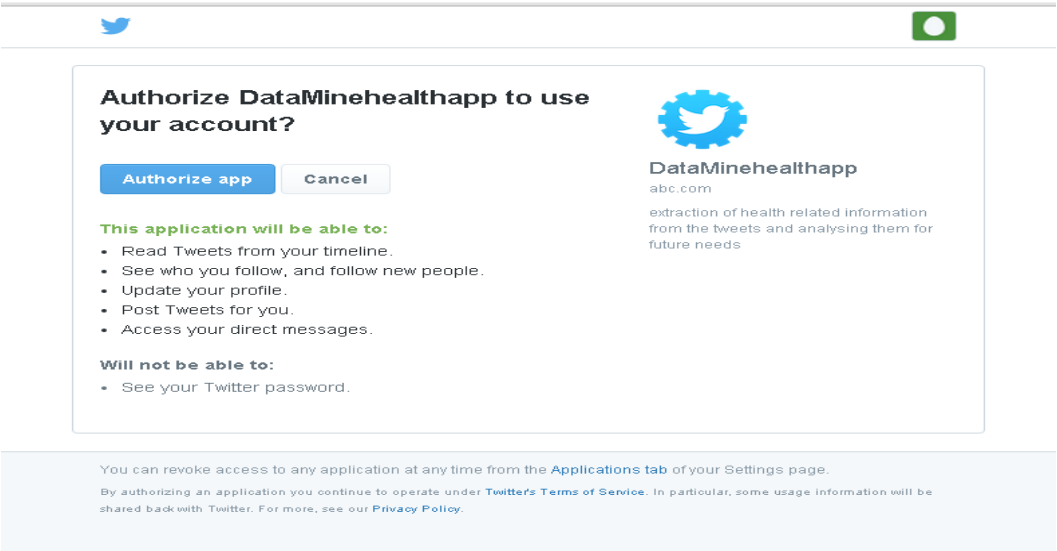

Figure 10: Authorization

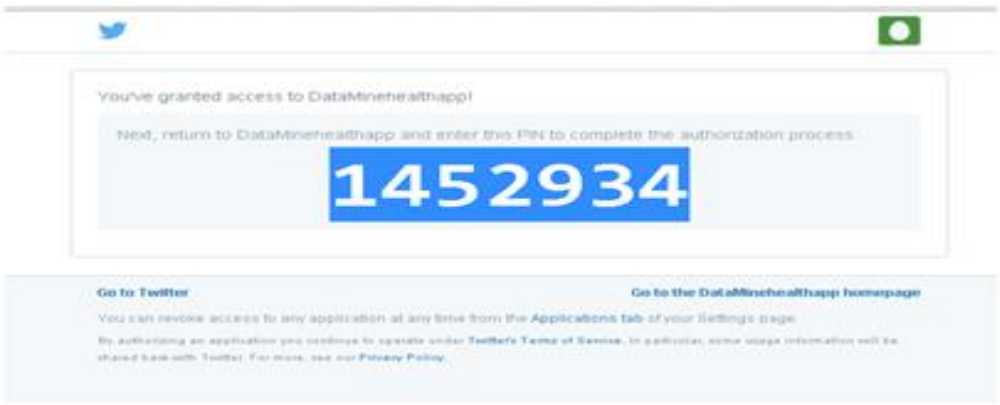

Figure 11:Pin for authentication

After the authorization and authentication process the user/developer has to specify the keyword in the console output (as shown in figure 12) so that only those tweets and its related information can be collected in which the keyword is present. Then once the specification of keyword is done then tweets get extracted on the console itself (figure 13) and other information is stored in database

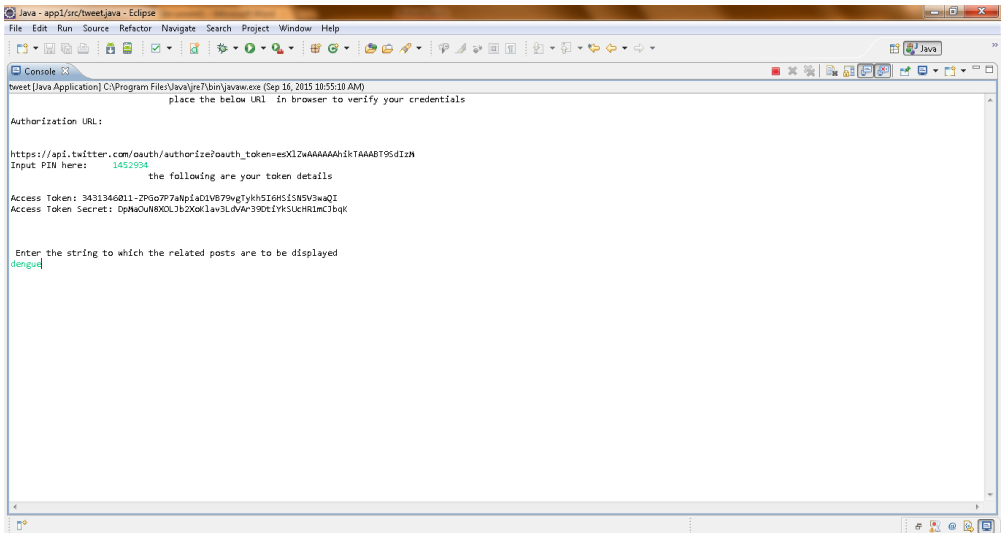

Figure 12: Keyword specified like dengue

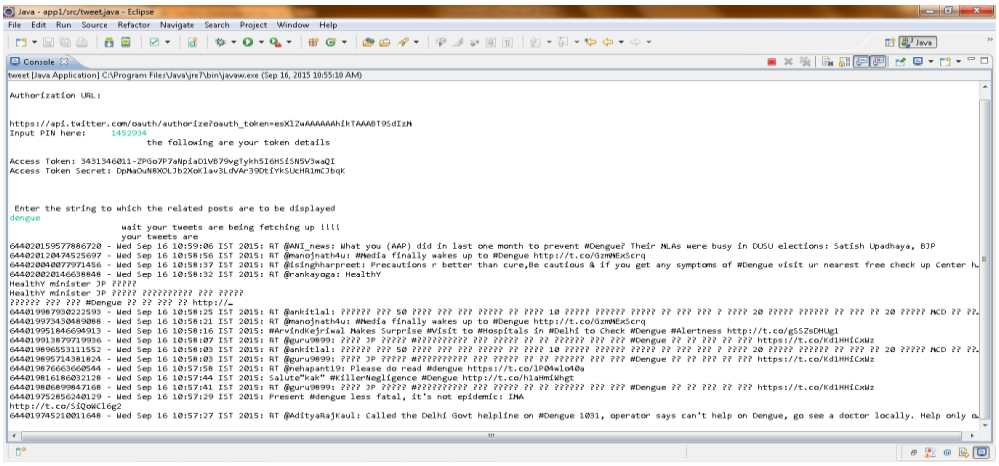

Figure 13:tweets extracted for Health analysis

3.2STORING THE EXTRACTED DATA

The extracted data from Twitter in stored into excel files (in case of news analysis) and in database and text files (in case of celebrity and health analysis).

3.2.1STORING THE TWITTER DATA IN EXCEL FILES

In news analysis, after extracting Tweets and other data from Twitter, it is then stored in excel files for further analysis and visualisation.

As shown in figure 14,the tweet posts and other data like usernames,retweet count,id,location,source are stored in excel files for further analysis and visualisation of news.

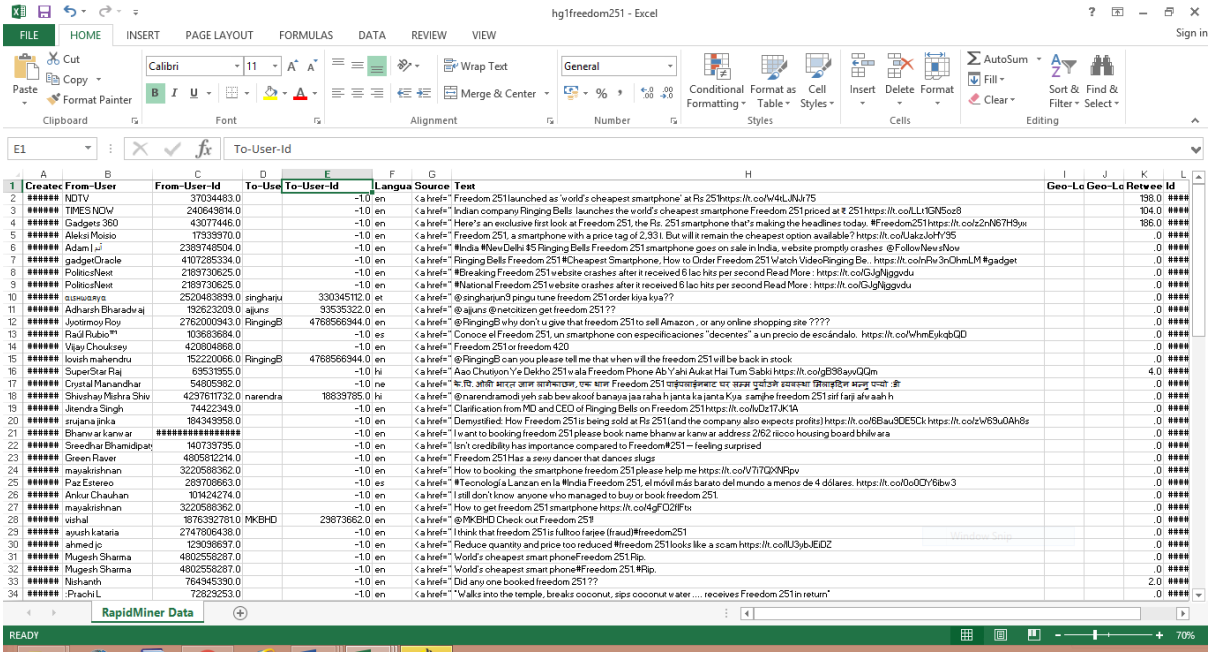

Figure 14: Excel file storage of news tweets

3.2.2STORING THE TWITTER DATA IN DATABASE AND TEXT FILES

After storing the extracted attributes in the database for celebrity and health, appropriate queries were used to calculate the frequency of tweets and other issues required for analysis.

For storing the data extracted, Jdbc odbc has been implemented using oracle 10g express edition.

Figure 15 explains that the java application has been integrated through JDBC ODBC bridge driver and a vendor library as (ODBC14.JAR).java application has been integrated through odbc driver with oracle 10 g express edition database.

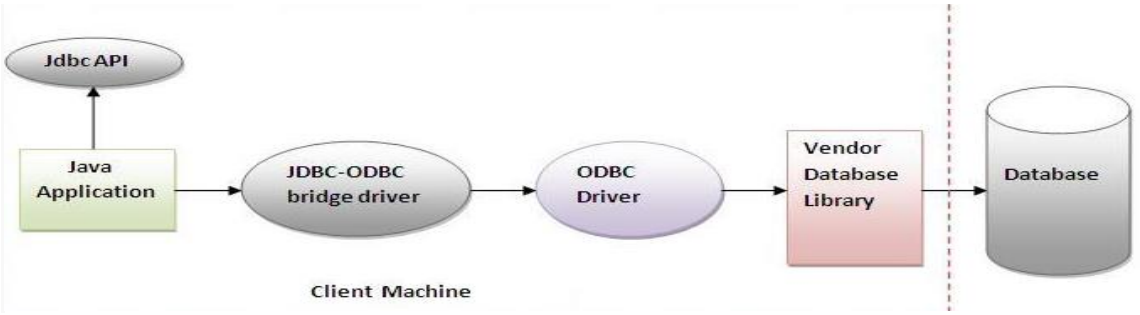

Figure 15: Client Machine

Storing the extracted data of celebrity

In this project we present a more general approach that analyze behaviour of any person from his/her tweets and behaviour of his/her followers towards him/her from tweets. This project first aims at extracting tweets from Twitter API. Various tweets' attributes like username, language of the tweet, month, year and location from where it was posted are extracted along with the tweet messages. These attributes, along with the original tweet message are then stored in a database. Two text files are stored which contains positive and negative words list. By comparing tweets of followers and other peoples ,tweets are analyzed to know behaviour about that person. For analyzing, Cirrus, another tool used for
analyzing the extracted data, is a word cloud displaying the frequency of words appearing in a corpus. This tool not only allows us to remove the stop words, but it also helps us to analyze the tweets frequency month wise to know activeness on twitter is increasing or decreasing .Using follower's geolocation, geomapping is done to find how much followers are lying in terror area so that it will help to find connection of that persons contacts and it will also help to analyse popularity of any celebrity varying country wise or any specific area wise. Two twitter data analysis tools are also analysed which are twitonomy and tweetchup to verify data which we get after java coding and by tool.

Tweets are extracted for celebrity " salman khan" and stored in DataBase

Firstly the tweets were extracted usingtwitter4j java library. Twitter4J is an unofficial Java library for the Twitter API. WithTwitter4J, we easily integrated our Java application with the Twitter service.

Firstly the tweets were filtered on the basis on language constraint in order to extract tweets where language of the follower is English using "en"constraint where "en" stands for English language. After the application of language constraints, the tweet related information such as username, his id number, language,timestamp, geolocation and his/her posted tweet were extracted. Then the tweets were saved in text file and other information that i.e. username, his id number, language,timestamp, geolocation were stored in database so that sql queries can be applied for analysis.
Figure 16represents the collection of tweets extracted of the celebrity "Salman Khan"
With the id "BeingSalmanKhan". Figure 17 represents the description of the table "celebrity". the screenname column represents the screenname or username of the follower of salman khan. language stores the language and location stores the location from where the user has tweeted and time represents their time of tweeting.

Figure 16:Tweets extracted for celebrity

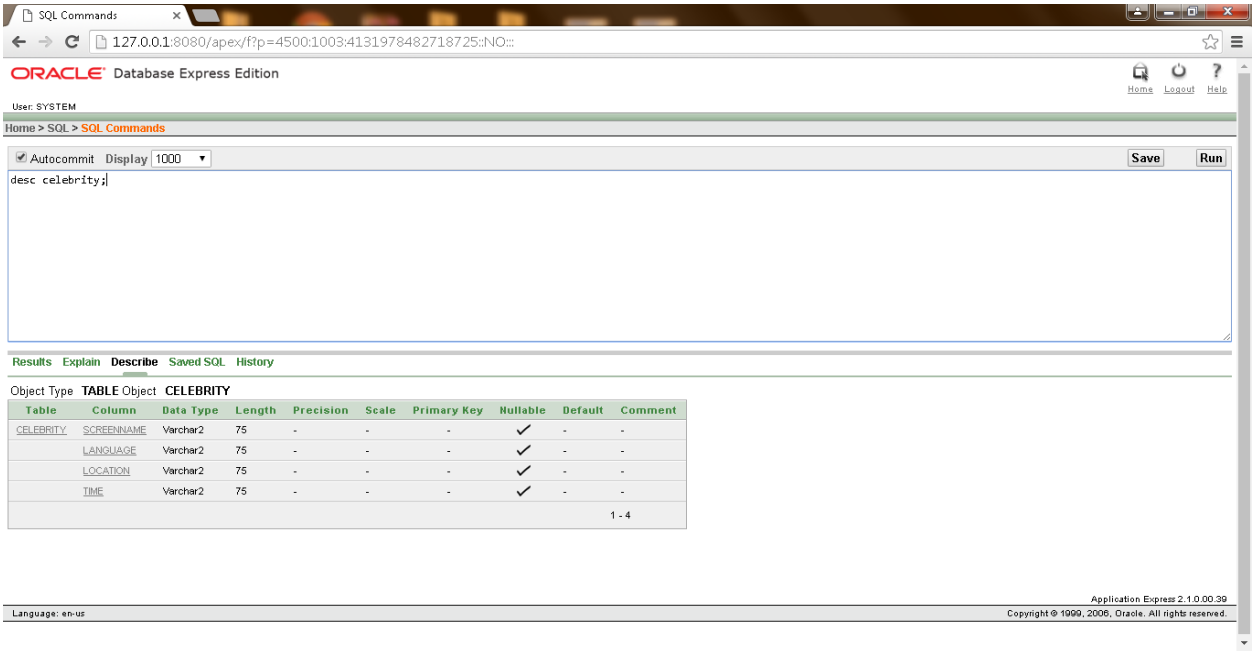

Figure 17:Tweets Storage in Database for celebrity

Categorization of tweets extracted into "positive" and "negative" on the basis of ontology specified

The tweets extracted are further analysed using the java code into positive and negative tweets on the basis of ontology as specified by us. firstly there are two text files containing positive and negative words which are fed as input to java code with the tweets of salman khan and as result various text files are created showing which the positive tweets are and why or due to which word it has been stated as positive and same with the negative case.

Figure 18 represent the various negative word in each line of text file for example "2 faced","abort". Figure 19represent the various positive word in each line of text file for example "happy","a+".

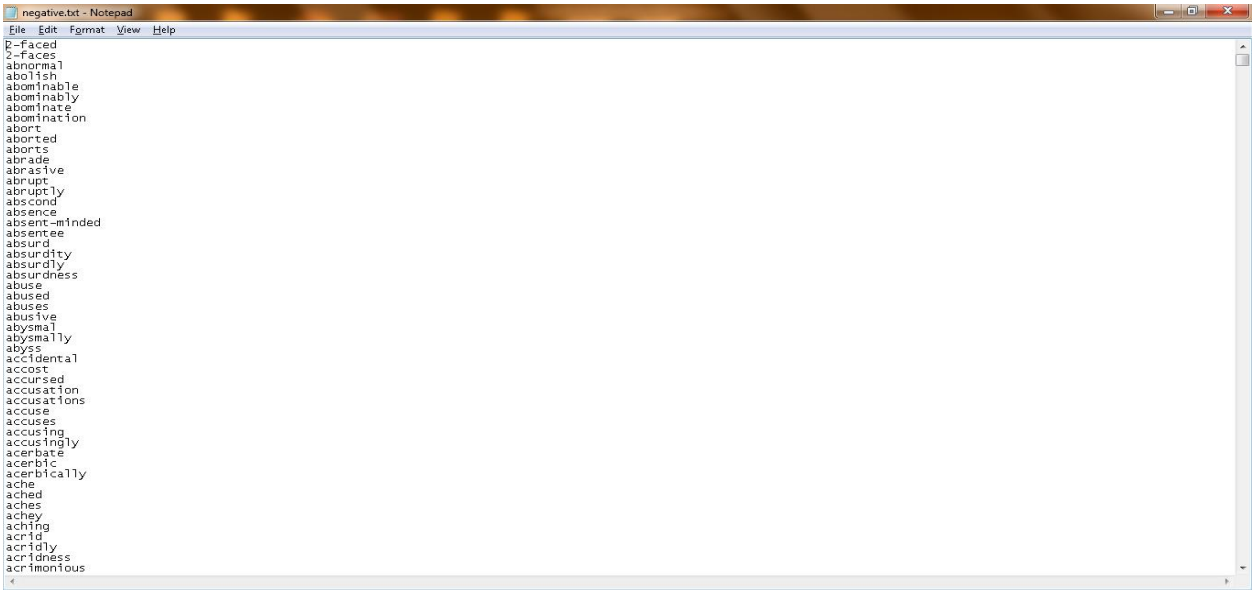

Figure 18:ontology of negative words

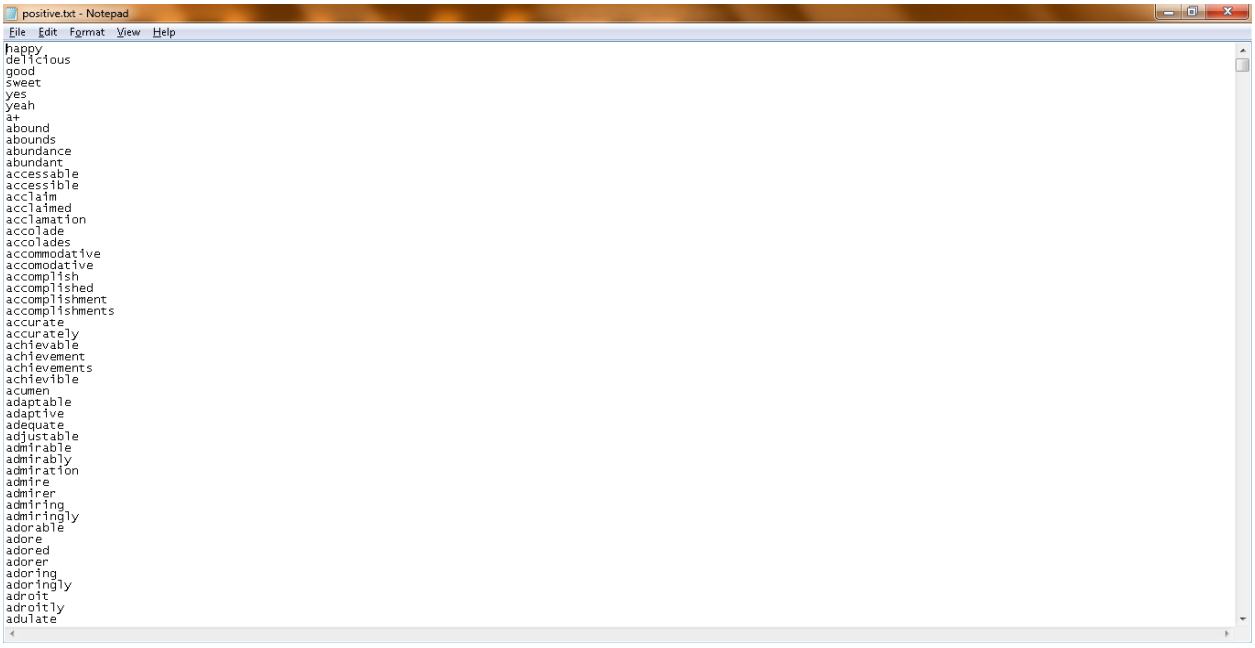

Figure 19:ontology of positive word

Figure 20: represents the categorization of positive tweets into text file from the text file containing both positive and negative tweets i.e salmankhan.txt and Figure 21: represents the positive words which have been used from the ontology during categorization.for example the tweet" First round of legal battle in #Salman case in SC won by Attorney General who convinces court that appeal needs to be exa…" is positive due to the word"won" stored in ontology.

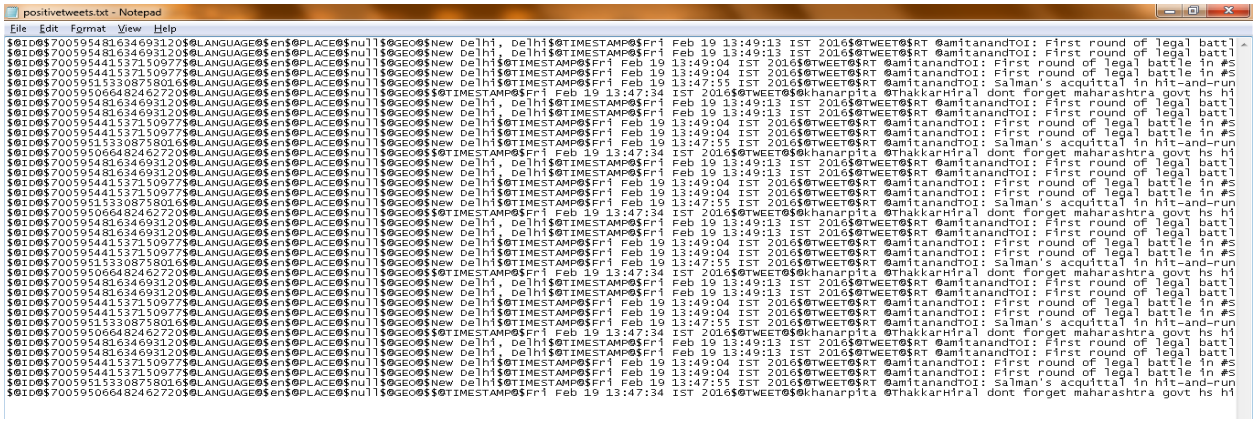

Figure 20:positive tweets extracted for keyword"BeingSalmanKhan"

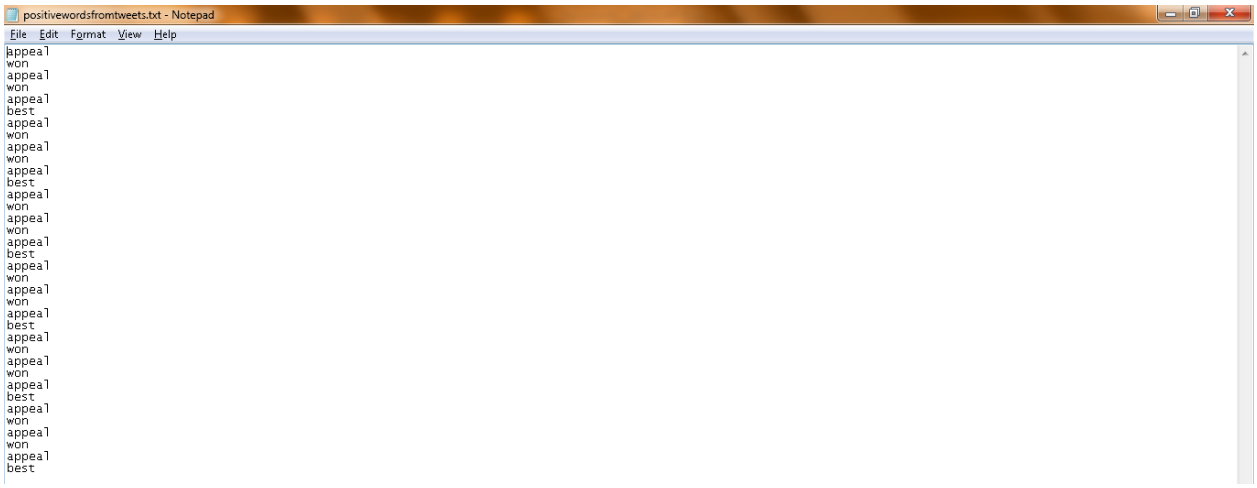

Figure 21:positive words used from the ontology

Fig 22: represents the categorization of negative tweets into text file from the text file containing both positive and negative tweets i.e salmankhan.txt and Fig 23: represents the negative words which have been used from the ontology during categorization.for example the tweet"SC sends notice to Salman on plea against acquittal" is negative due to the word"plea" stored in ontology.

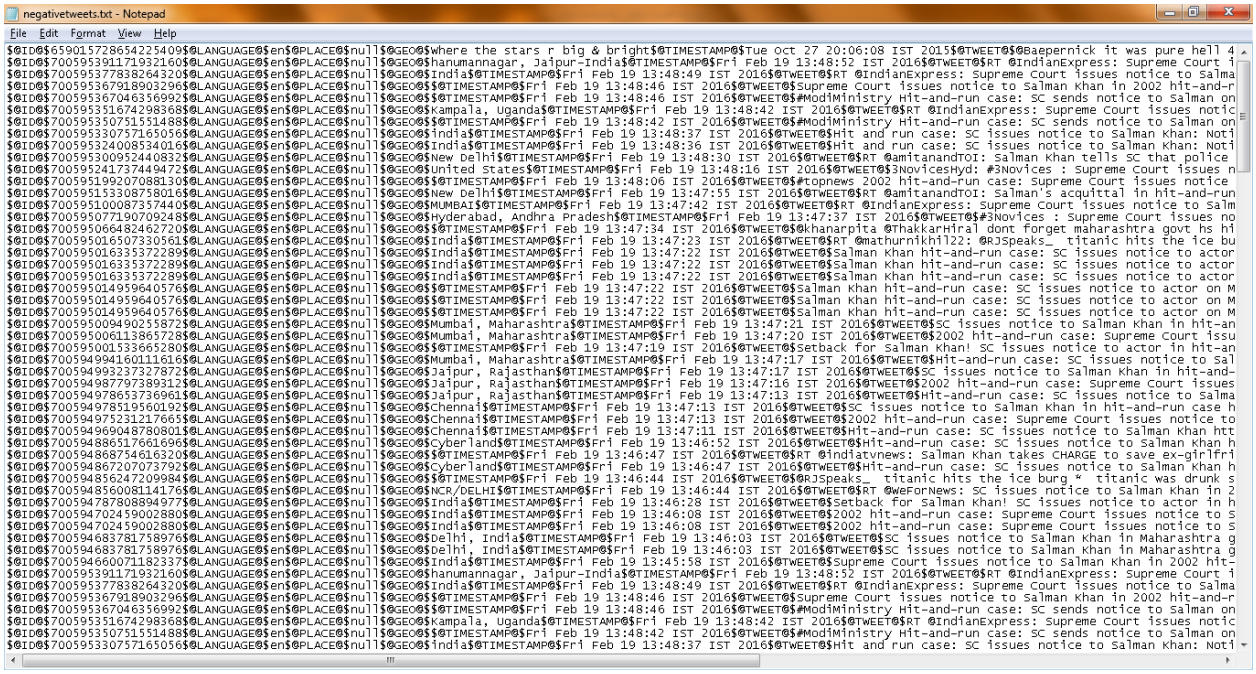

Figure 22:negative tweets extracted for keyword"BeingSalmanKhan

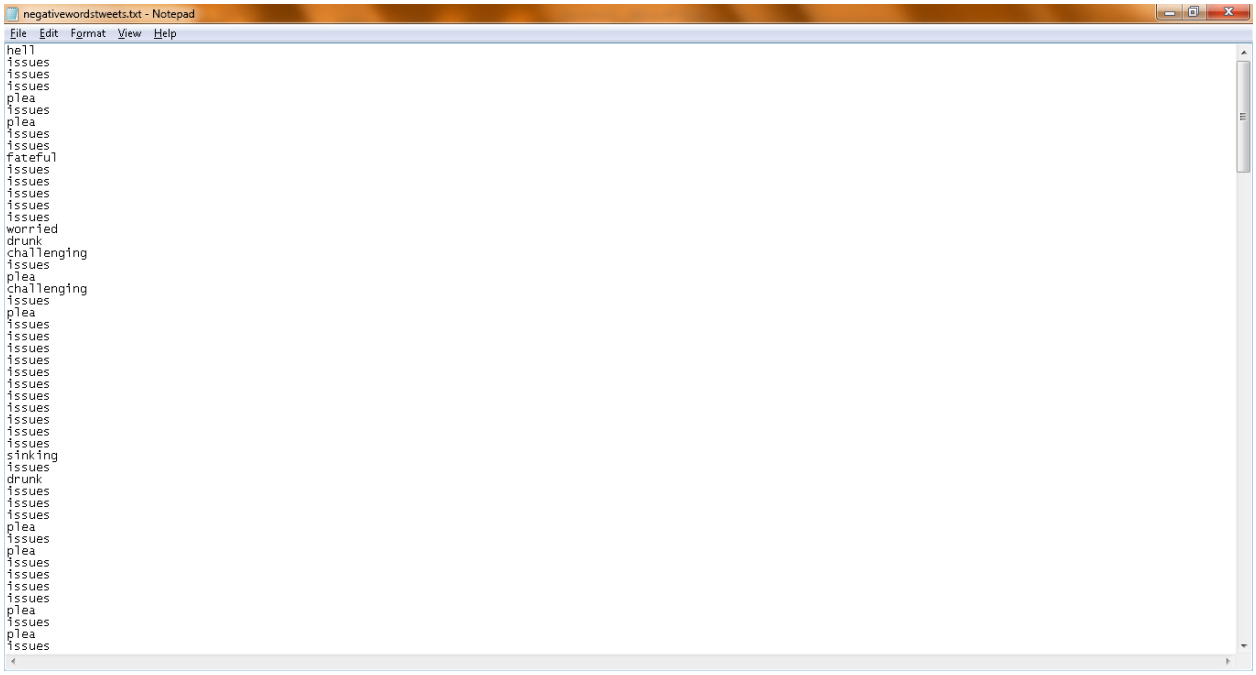

Figure 23:negative words used from the ontology

Follower details for salman khan are extracted and stored in text file

The followers of salman khan on twitter are extracted and their details are stored in database.

Fig.24 contains two column screenname representing the follower name with their location.

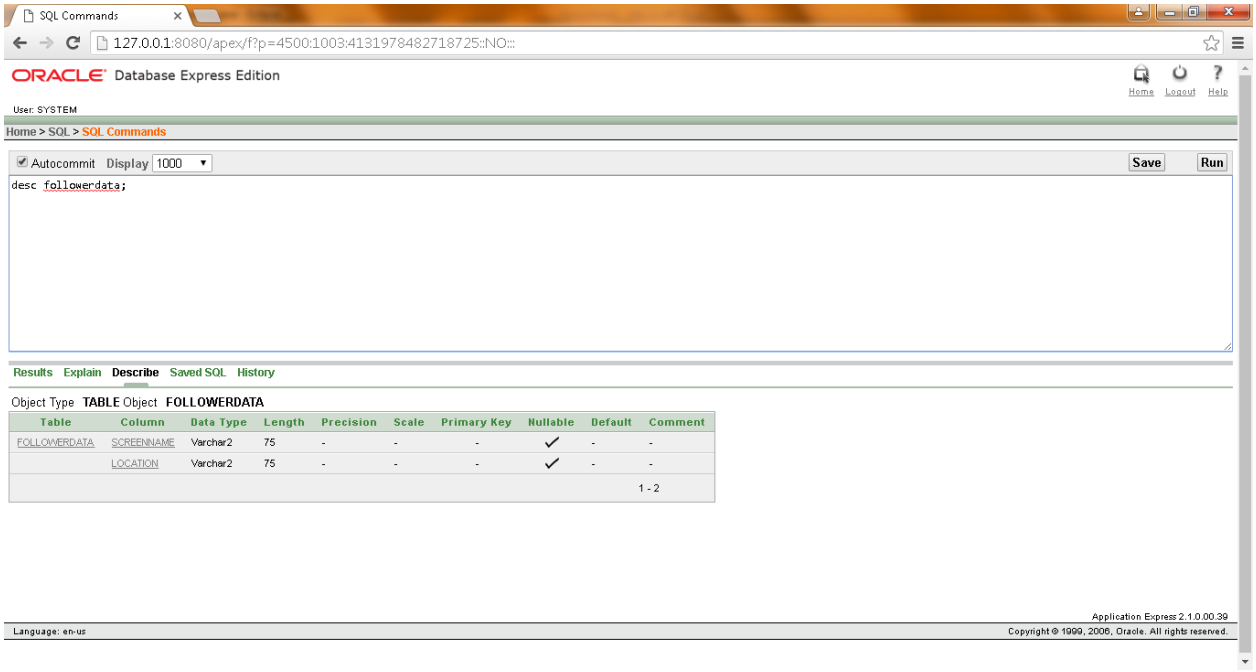

Figure 24:database for followerData

The location extracted is then analysed using geolocation showing which country has more followers and terror stated like khazakistan does has any follower indicating the popularity of the celebrity region wise

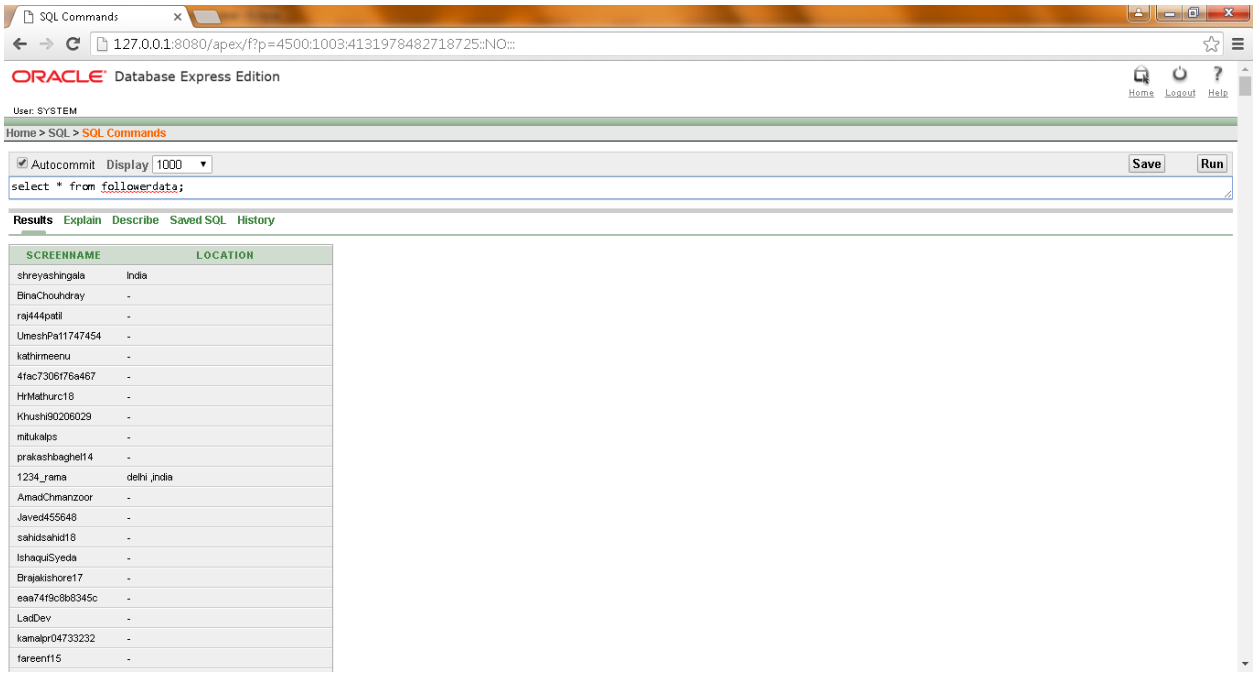

Figure 25:follower detail with their location

Fig.25 shows the relative position of various followers from various parts of country.for example “shreyashingala is from”india”

This data can now be exported to csv file and then later on fed to the Maptive to generate geolocations showing the user and the relative frequency of the location.further. it can be categorized into country wise mapping which indicates which country is active

Follower views about salman khan are extracted and analysed monthwise

The views of the salman khan followers are extracted using java code month wise by paging.the count of relative month has been collected which can further be used to indicate as to which month did the salman khan was popular among people or their followers and by what percentage

Fig26,27,28:shows the collection of follower views in various months like September(sep.txt) December January(jan.txt) and which are further counted using java code

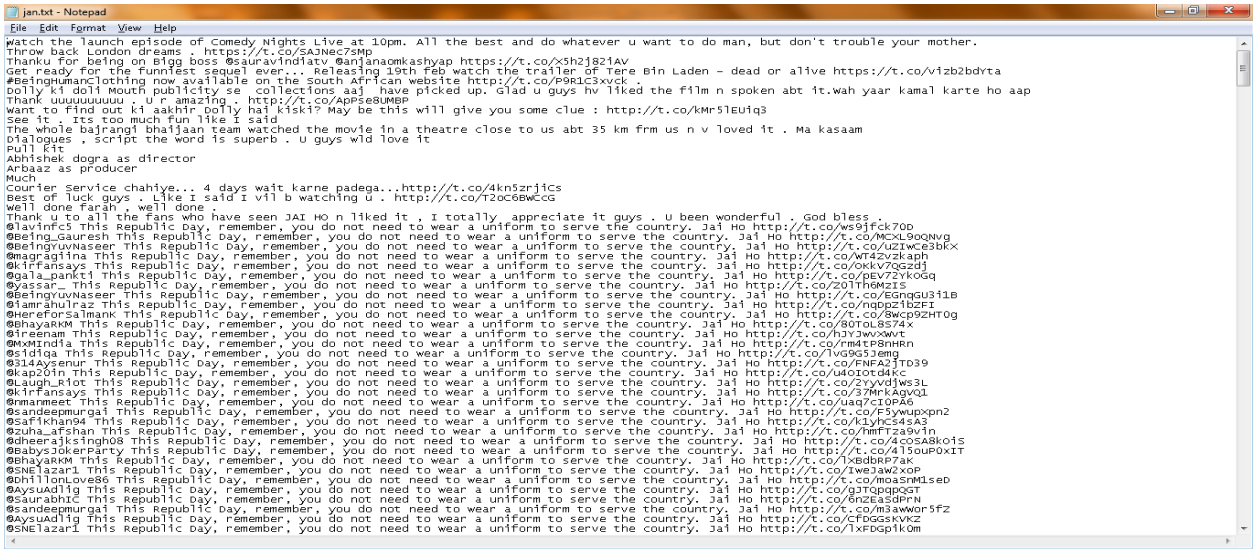

Figure 26: follower views in January

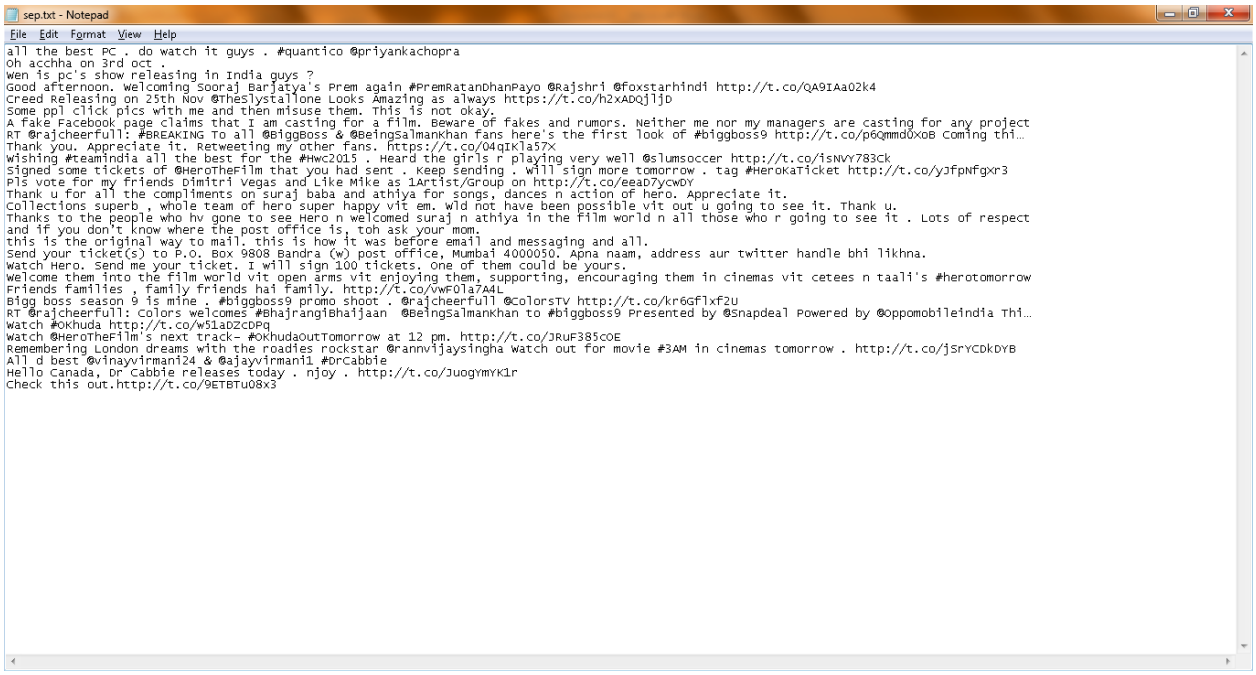

Figure 27: follower views in September

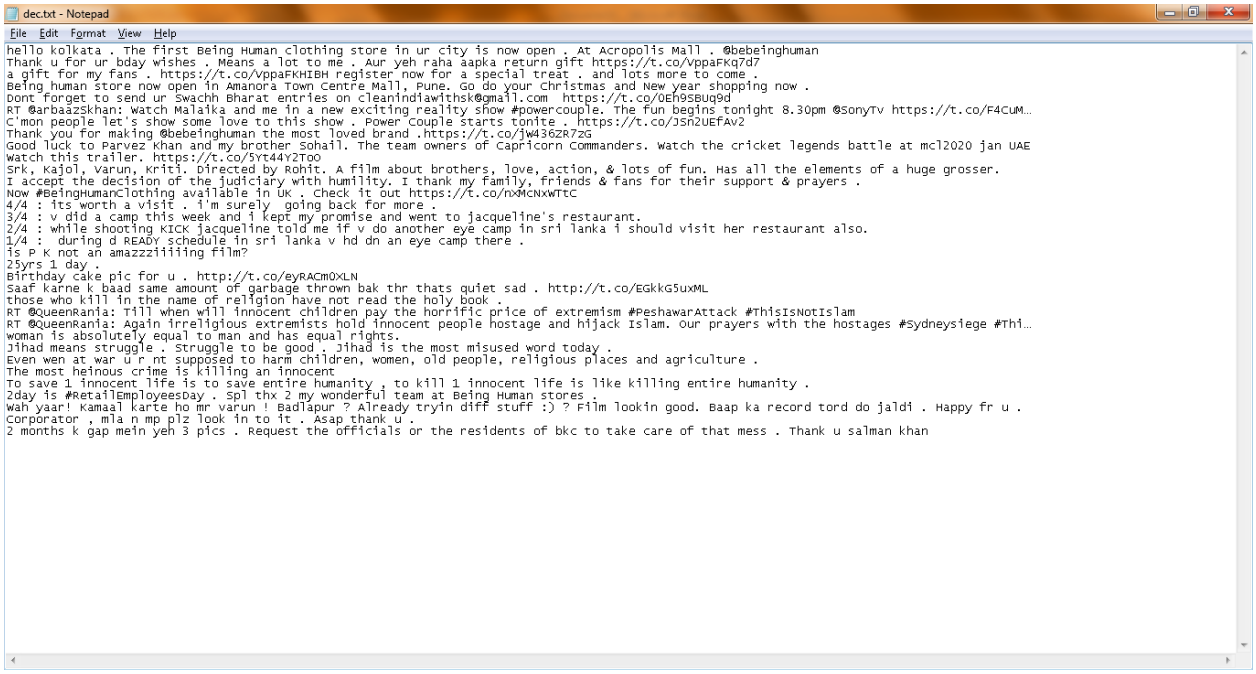

Figure 28: follower views in December

Fig 29:indicated the relative count of tweets of followers in various months.it indicates that tweet count is highest in January(646) and lowest in May(9).

The count of these months can now be plotted on plotly

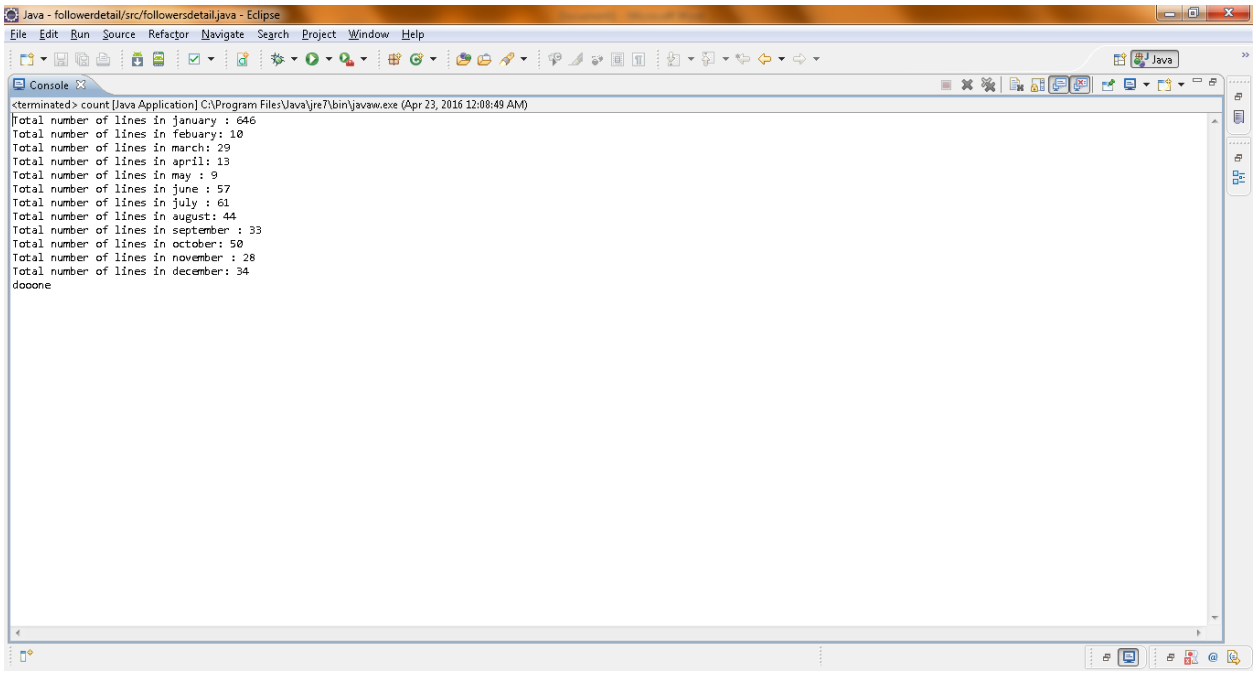

Figure 29:count of follower views in various months.

Storing the health related data in database and text files

Figure 30 describes the table storing the attributes of tweets extracted. The table is named twittergeo storing the user id, language in which tweets has been extracted, which in this case has been restricted to en (English), this merely acts a constraint. The next column is of geo which specifies the location of user from where it has been posted and year describes timestamp of the tweets extracted. Figure 31 is the magnified view to the twittergeo table shown in figure 30

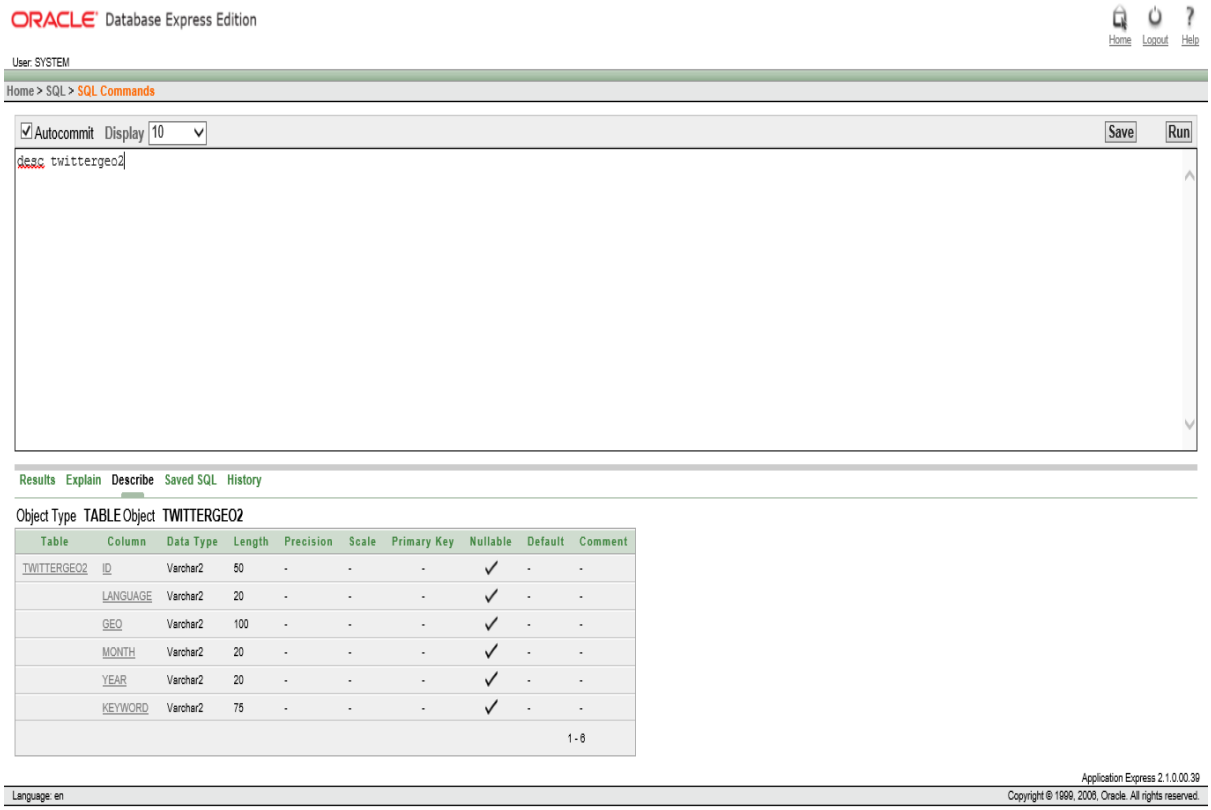

Figure 30 :description of table storing the tweets the data type of entities.

Table	Column	Data Type	Length	Precision	Scale	Primary Key	Nullable	Default	Comment
TWITTERGEO2	ID	Varchar2	50	-	-	-	✓	-	-
	LANGUAGE	Varchar2	20	-	-	-	✓	-	-
	GEO	Varchar2	100	-	-	-	✓	-	-
	MONTH	Varchar2	20	-	-	-	✓	-	-
	YEAR	Varchar2	20	-	-	-	✓	-	-
	KEYWORD	Varchar2	75	-	-	-	✓	-	-

1 - 6

Figure 31: detailed description

Figure 32 reflects the data stored in twitter geo by selecting the rows from the table. For instance the first row describes the id and the language as en, geo as i=United Kingdom month as October year as 2015 and keyword cancer. Figure 33 is the magnified view of the result obtained by selecting the rows of twittergeo.Figure 34 shows the text file storage of tweets for cloud visualisation analysis.

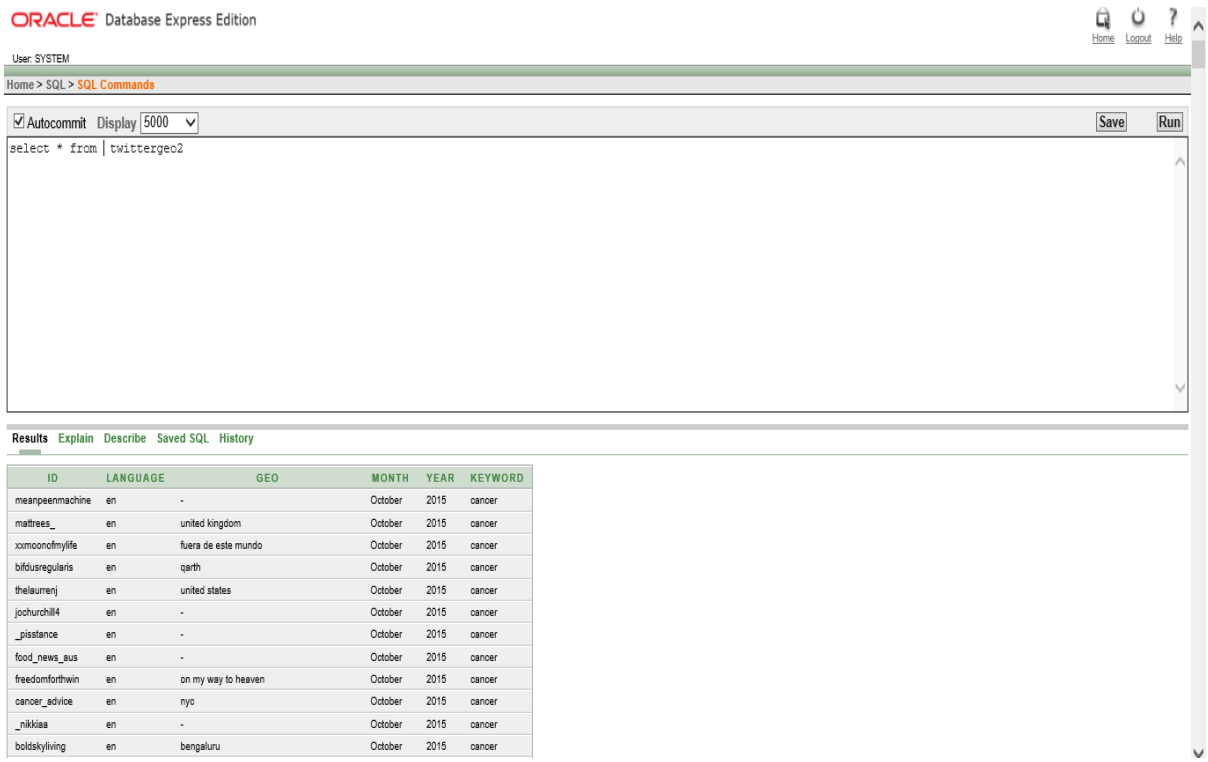

Figure 32: stored data in table twittergeo

ID	LANGUAGE	GEO	MONTH	YEAR	KEYWORD
meanpeenmachine	en	-	October	2015	cancer
mattrees_	en	united kingdom	October	2015	cancer
xxmoonofmylife	en	fuera de este mundo	October	2015	cancer
bifdusregularis	en	qarth	October	2015	cancer
thelaurrenj	en	united states	October	2015	cancer
jochurchill4	en	-	October	2015	cancer
_pisstance	en	-	October	2015	cancer
food_news_aus	en	-	October	2015	cancer
freedomforthwin	en	on my way to heaven	October	2015	cancer
cancer_advice	en	nyc	October	2015	cancer
_nikkiaa	en	-	October	2015	cancer
boldskyliving	en	bengaluru	October	2015	cancer
amieynnn	en	peninsula ✈ borneo	October	2015	cancer
homaihass	en	johor darul takzim	October	2015	cancer
tenaj_gerundio	en	cebu	October	2015	cancer
mrkeitheb	en	-	October	2015	cancer
navosnavos476	en	-	October	2015	cancer
so_tsuda	en	glasgow, scotland	October	2015	cancer
wilfreymorena	en	-	October	2015	cancer

Figure 33: detailed data stored in twittergeo

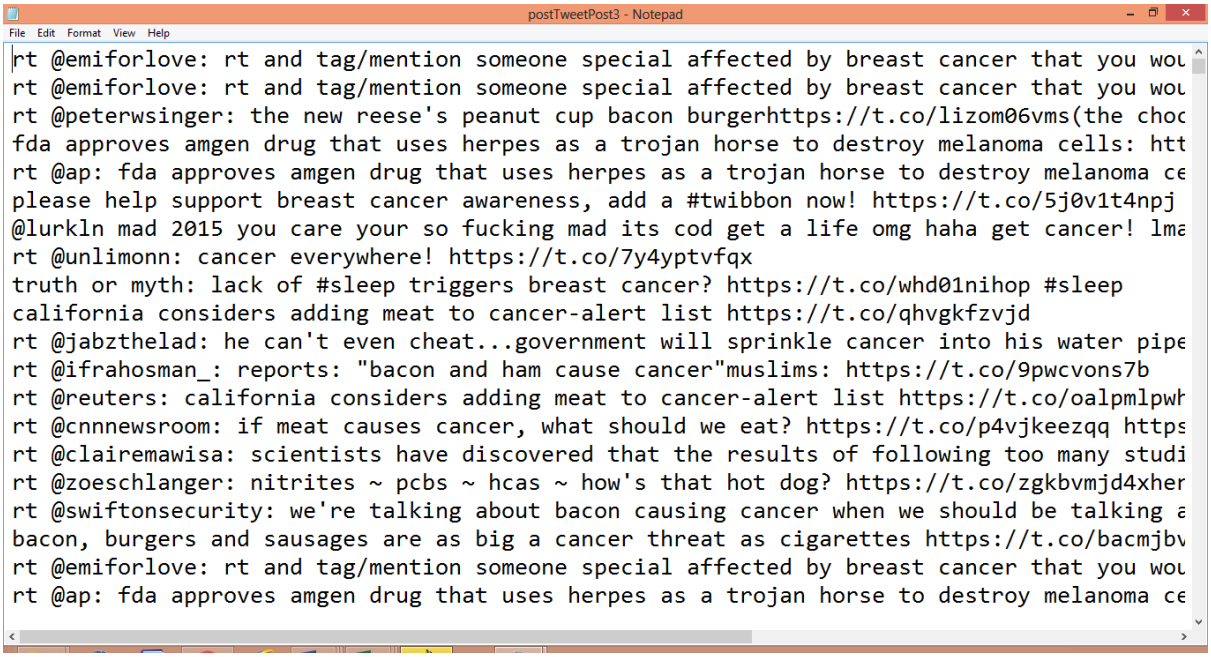

Figure 34: text file storage for tweets

CHAPTER 4

REPRESENTATION AND ANALYSIS TOOLS

4.1 Graphical Analysis

The graphical analysis of Twitter Data Analytics is done to conclude results made on particular topic as per the 3 categories kept under consideration.

Figure 35 and 36 shows the process and output of tweet analysis in RapidMiner using Aylien Extension.The output gives the polarity and subjectivity confidence values.

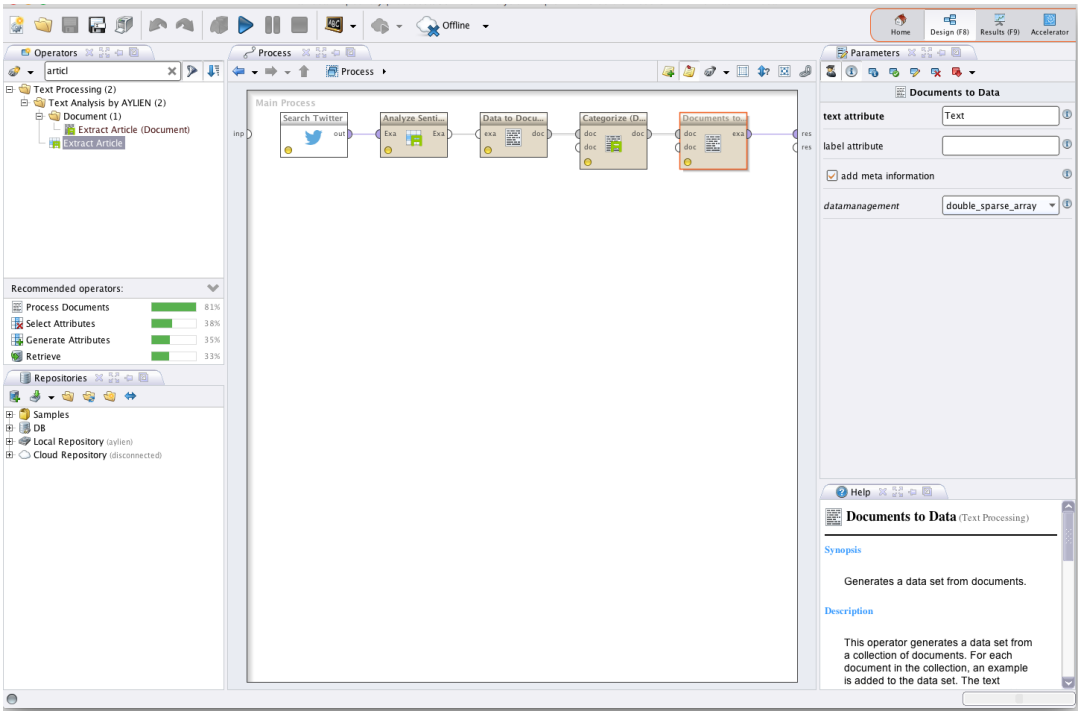

Figure 35: Process in rapidminer for analysing tweets

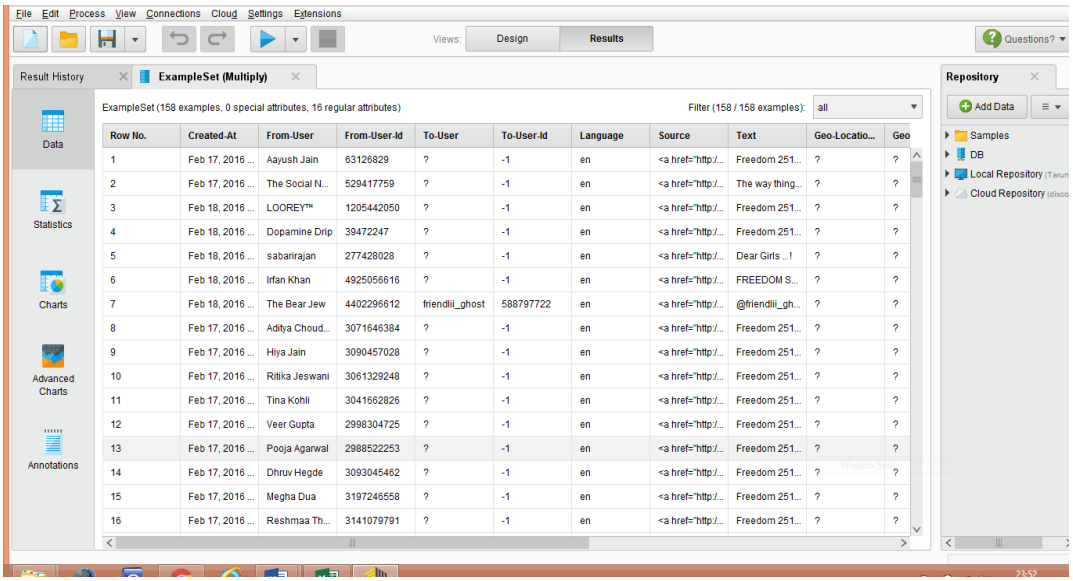

Figure 36: output on tweet analysis

4.1.1News Analysis

The graphs for this shows many valuable results along with some proofs.

1. **Polarity analysis** to show whether a particular news is positive,negative or neutral by showing the count & from that we can infer the thoughts of people.This analysis is based on ontological criteria of Aylien extension.
   As in figure 37, count of polarity shows that people have neutral feeling(113 neutral tweets out of 158) about Freedom 251,hence they don't possess any liking or disliking to freedom 251 launch.

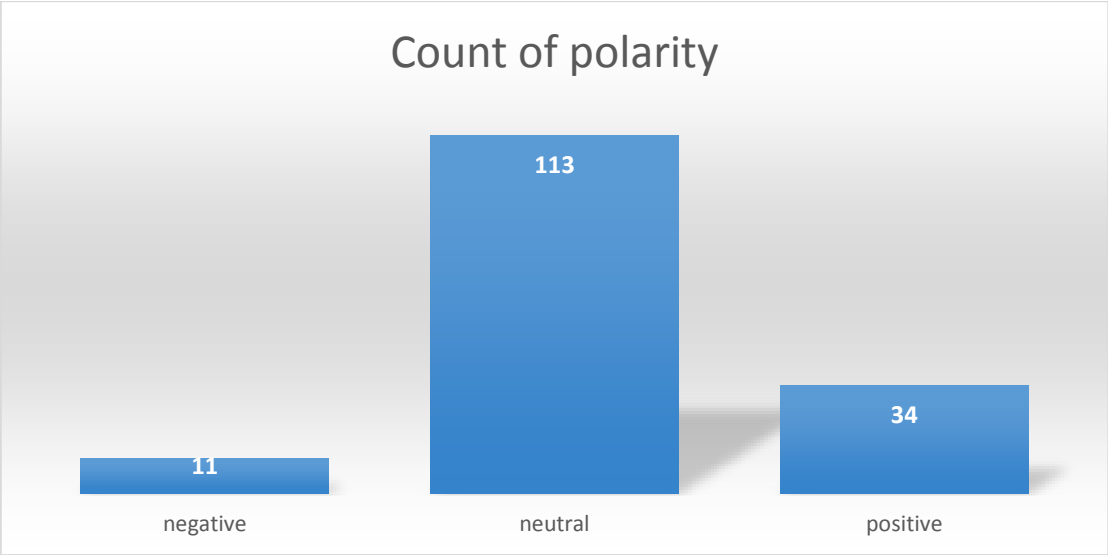

Figure 37: count of polarity

Proof:

I. **method1** :applying ensemble technique-"vote"
   Here the data mining techniques which are ensemble are decision tree, neural networks and LDA

Figure 38 shows the output of the decision tree interpreting the result as neutral. The output of LDA is shown in figure 39 infering the support that freedom251 is a neutral news.The figure 40 shows the output of neural networks where the input is polarity_confidence and subjectivity_confidence and output nodes are positive,negative and neutral.The figure 41 shows the stacking model of the vote ensemble technique and the result shows the neutral support of the public regarding feedom251.

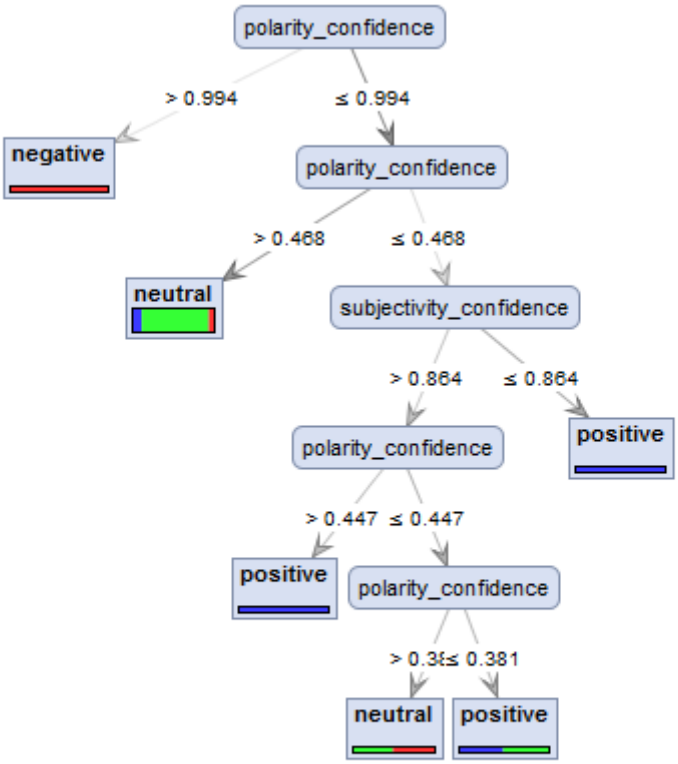

Figure 38: Output of decision tree

Quadratic Discriminant Model

```
Apriori probabilities:
positive        0.2152
neutral 0.7152
negative        0.0696
```

Figure 39: Output of LDA

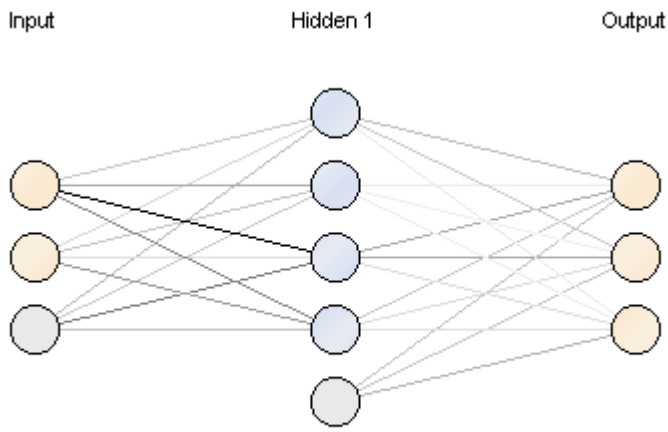

Figure 40: Output of Neural Networks

AttributeBasedVoting

```
Using the majority of the following attributes for prediction:
  base_prediction0
  base_prediction1
  base_prediction2

The default value is neutral
```

Figure 41: Overall output of vote ensemble technique

II. **method2**: retweet frequency count: through which it can be inferred that which particular tweet is tweeted highest and supports which kind of polarity.
The figure 42 shows the highest counts for neutral polarity and hence again proves the neutrality of freedom251 news.

Text	Retweet-Cou	polarity
Freedom 251 launched as 'world's cheapest smartphone' at Rs 251https://t.co/W4tLJNJr75	198.0	neutral
Here's an exclusive first look at Freedom 251, the Rs. 251 smartphone that's making the headlines today. #Freedom251 https://t.co/	186.0	neutral
Indian company Ringing Bells launches the world's cheapest smartphone Freedom 251 priced at ₹ 251 https://t.co/LLt1GN5oz8	104.0	neutral
Tommorow one new Indian company Freedom is launching Lollipop Smartphone with quad core processor at Rs.251. This is Jalwa of	8.0	neutral
251₹ की कीमत के साथ FreeDom कम्पनी ने लॉन्च किया बेहतरीन फीचर्स वाला Android स्मार्टफोन.....https://t.co/hRPTCQGBFs https://t.	5.0	neutral
Aao Chutiyon Ye Dekho 251 wala Freedom Phone Ab Yahi Aukat Hai Tum Sabki https://t.co/gB98ayvQQm	4.0	neutral
Did any one booked freedom 251 ??	2.0	neutral
Freedom 251!!!!!	1.0	positive
The way things are going, I'm pretty sure tomorrow morning at 6 AM the freedom 251 website will crash ?? ; Rs 251 @makeinindia	1.0	positive
రు.251లకే స్మార్ట్ ఫోన్! ఆవిష్కరించనున్న కేంద్రమంత్రిhttps://t.co/JrLOFJPw2Z #Freedom251 https://t.co/wuwYOx5Pi5	1.0	neutral
के.पि. ओली भारत जान लागेकाछन, एक थान Freedom 251 पाईपलाईनबाट घर सम्म पुर्याउने व्यवस्था मिलाइदिन भन्नु पऱ्यो :डी	1.0	neutral
I'm getting badly trolled on Twitter and Facebook because I called out the Freedom 251 phone and its Made in India claims. This is hc	1.0	negative
Freedom 251. Smartphone at Rs. 251. Pretty cool. World's cheapest phone ever.	.0	positive
Freedom 251 seems to be the Akash Tab of smartphones!	.0	positive
Dear Girls .. !!# Marry that Boy who was successfully Able to Place the Order of Freedom 251 .. !! ;) ;) # Patience Level _/_	.0	positive
FREEDOM SMART PHONE AT@251 Then 3G Recharge @300 +????????????	0	positive

Figure 42: retweet count of freedom 251 posts

2. **Subjectivity analysis**to analyse the public opinion using object keywords like "RINGING BELLS","FREEDOM251","IPHONE","600000 REQUESTS "ETC. to discuss the generic v/s particular news support of the public. This analysis is based on ontological criteria of Aylien extension.
   Figure 43 shows that out of 158 tweets,85 are objective and 73 are subjective.

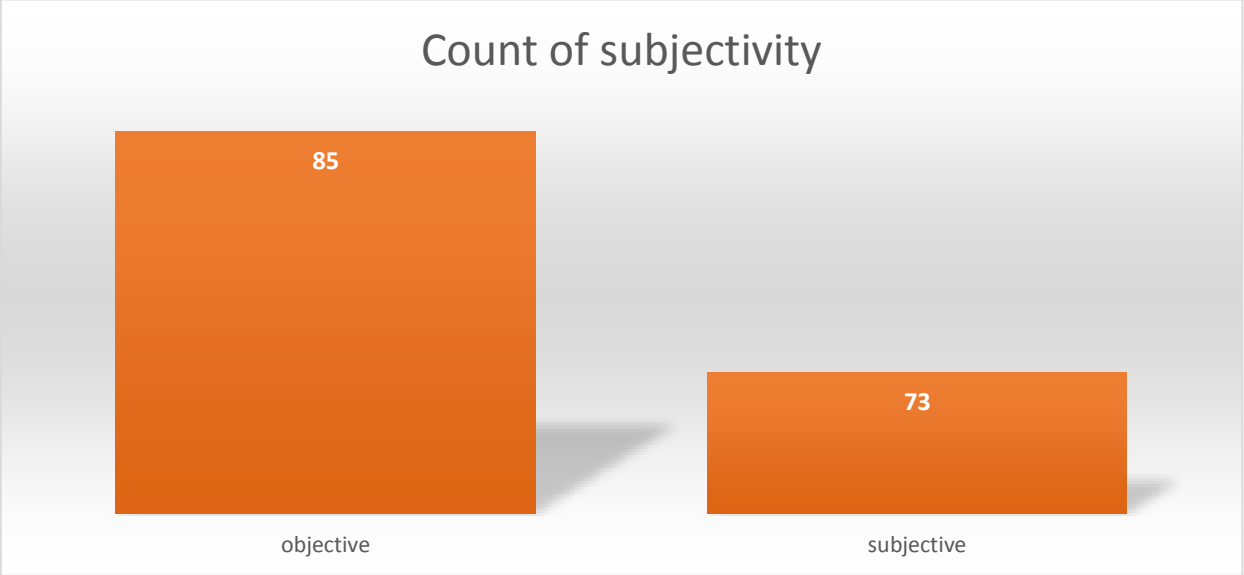

Figure 43: count of subjectivity

3. **ACCURACY ANALYSIS OF POLARITY** BY COMPARING ACCURACY RATE OF DECISION TREE, NAÏVE BAYES AND K-NN
   The figure 44 shows that decision tree(77.33%) gives the highest accuracy in determining the neutral support of the tweets

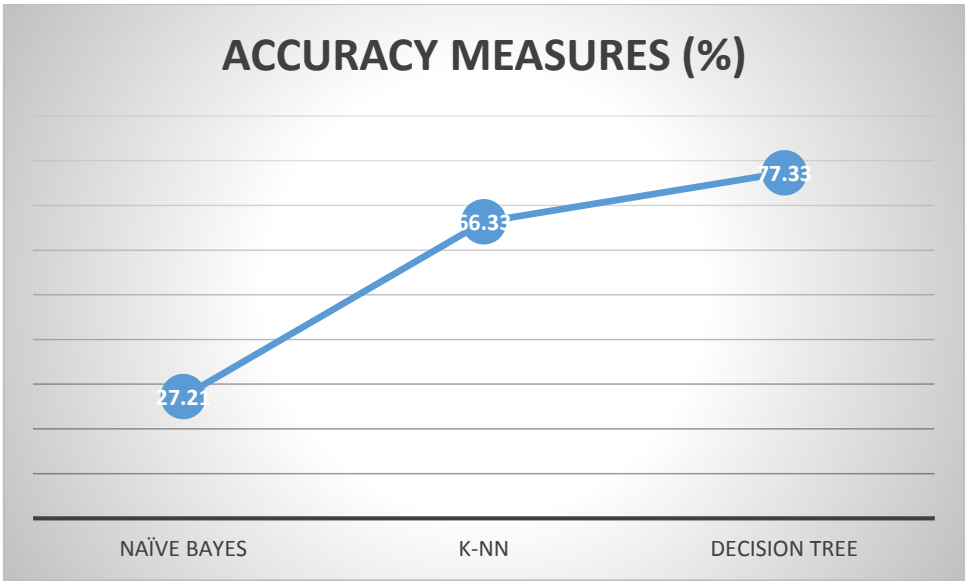

Figure 44: % accuracy comparison of data mining algorithms for polarity confidence

4. **ACCURACY ANALYSIS OF SUBJECTIVITY** BY COMPARING ACCURACY RATE OF DECISION TREE, NAÏVE BAYES AND K-NN
   The figure 45 shows that decision tree(57.67%) gives the highest accuracy in determining the subjective nature of the tweets

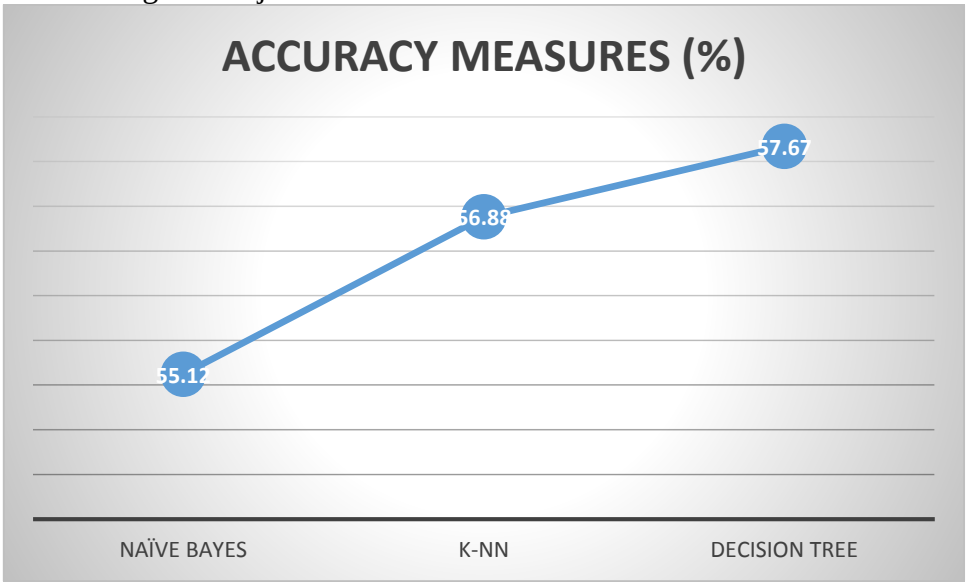

Figure 45: % accuracy comparison of data mining algorithms for subjectivity confidence

4.1.2Celebrity Analysis

The tweet count of various followers in various months have been collected using java code which can be plotted using plotly.

Fig. 46 show that it was highest in January and steep fall in febuary which later on enhanced a little bit in march with sudden decrease in may and the almost a near about after july. Fig. 47 is an another representation of distributive which exactly displays the

relative amount by which the tweet count has been changing in various months i.e maximum in January and minimum in may.

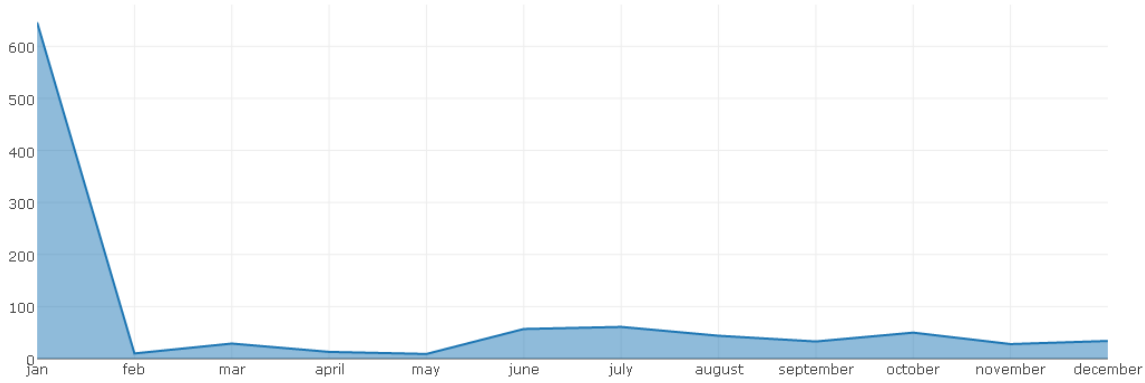

Figure 46:Distributive analysis of months vs follower tweet count

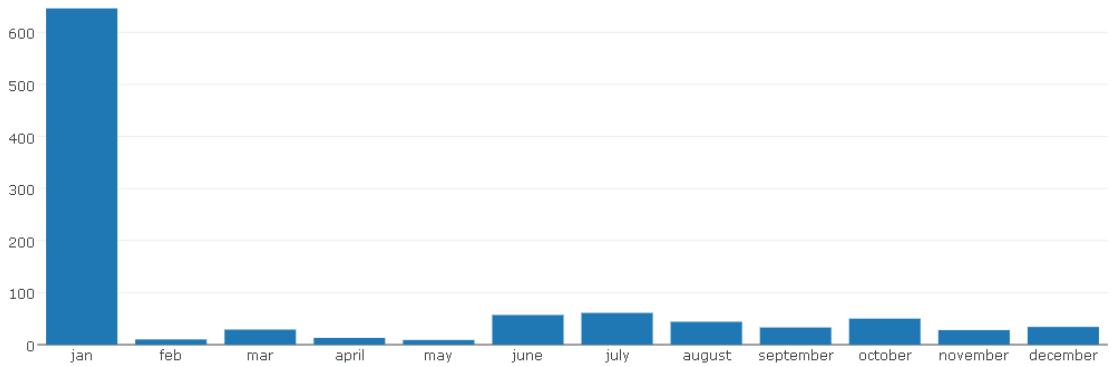

Figure 47:Bar Graph for month vs tweet count

The graphical analysis further yield JSON code which can be used in future to build up a web service for online creating of graphs

- {
  - data:[
    - {
      - name:"MFEB",
      - type:"bar",
      - xsrc:"tanyabansal:3:DOBWYZIY0J3B4A0GDBEQTSLEF000EW86",
      - ysrc:"tanyabansal:3:V43NE28ZRHDFOJJY8CR991WX41QXYYDU",
      - uid:"dd6280"

```
                }

        ],
        layout:{
            yaxis:{
                title:"",
                type:"linear",
                range:[
                    0,
                    680

                ],

                autorange:true

            },

            xaxis:{
                title:"",
                type:"category",
                range:[
                    -0.5,
                    11.5

                ],

                autorange:true

            },

            height:497,
            width:1121,
            autosize:true,
            dragmode:"zoom"

        }

}
```

4.1.3Health Analysis

After storing the extracted attributes in the database, appropriate queries were used to calculate the frequency of tweets related to a particular disease for each month. This gave us an idea about the months in which the number of cases of a particular disease was prevalent. Also, seasons and months having chronic and acute disease cases can be later analysed to prevent such outbreak by taking appropriate measures beforehand.

The graphical analysis shows the frequency of tweets posted related to a particular keyword with respect to the months. Here in figure 48 frequency of cancer related posts was considered such that the cancer posts in the month of November (the blue part in pie

chart) were 60 and that in October (the orange part in pie chart) were 75, which means that the followers were more concerned about cancer in October than in November .also here in the fig48 only two months were considered since the tweets were collected in these two months only.

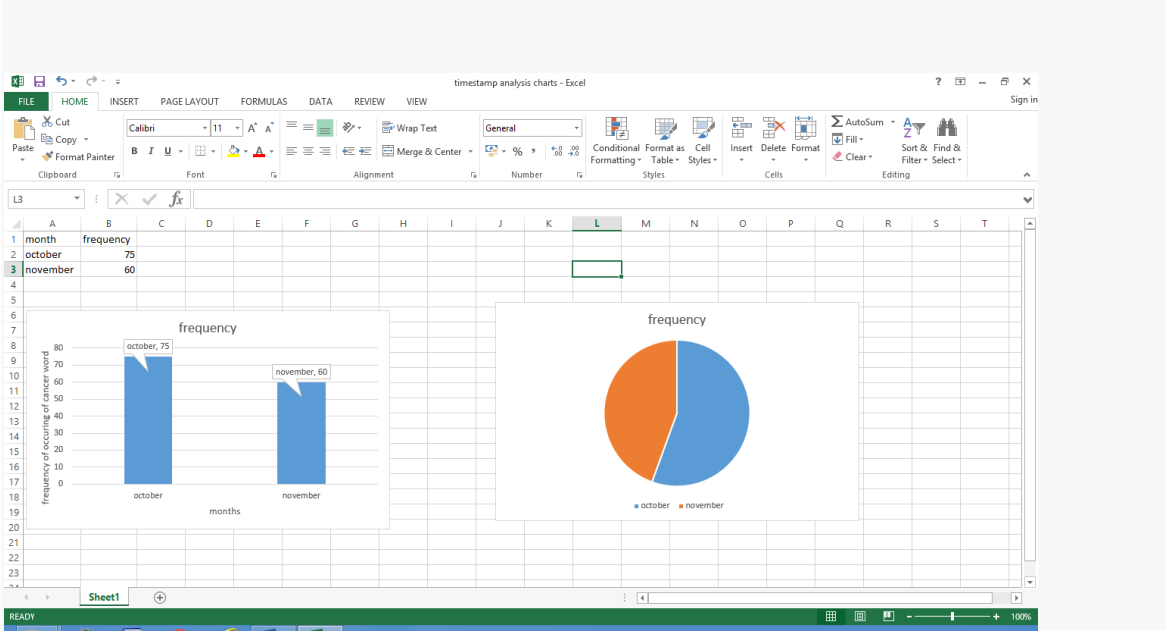

Figure 48: Frequency of tweets

4.2 Word Frequency Visualisation

Finally, we used cirrus tool to particularly analyse the data of tweet posts and extract the best known information from that. Cirrus is a visualization tool that displays a word cloud relating to the frequency of words appearing in one or more documents. One can click on any word appearing in the cloud to obtain detailed information about its relativity. Cirrus is a freely available visualization tool that generates a word cloud from a document or group of documents. The initial word cloud contains all of the most frequent words from the text, laid out graphically to fit together within an elliptical shape. Words may be oriented either vertically or horizontally, and they are rendered in different colours for ease of viewing. The user should be aware that word orientation and colour are strictly decorative elements, they do not indicate anything meaningful about the word. The size of the word, however, indicates the frequency with which it appears in the document. The larger the font size, the more frequent the word. The user can mouse over a word to see the precise number of times that it appears in the document.

One of the problems with word frequency visualization is the prevalence of conjunctions and articles in most texts. These less semantically interesting words often drown out the adjectives, adverbs, nouns, and verbs that are more likely to be of interest to the researcher. Cirrus offers an easy solution to this problem. By clicking the options icon in the upper right hand menu, a user can apply a list of stop words to the visualization. These stop words will be treated as noise, and stripped from the cloud, allowing words of greater interest to surface. Cirrus offers two pre-built lists of stop words, one for common English words, and one for common French words. The user can view the lists, so she know which words will be removed before applying them.

Cirrus offers users the ability to search within a word cloud for specific words, which is especially useful for seeing all of the word variations in a text. If I enter “comput” in the search box, Cirrus returns a word cloud indicating the word frequency of related terms like computers, computational, computing, computer-generated, computer-literate, microcomputers, human-computer, and supercomputing. It is also possible to enter comma separated terms in the search box in order to compare the frequency of multiple terms. By entering “comput, digit” for example, the user sees all of the variations of computer, as well as digital, digitization, etc. in the resulting word cloud. Any number of terms can be compared in this way. The Cirrus export menu exposes several options for sharing a word cloud. A permanent URL can be generated for a particular visualization, but the software also provides code snippets that allow the word cloud to be embedded in a website as a button or as a full visualization. Using this tool we basically analysed the certain important keywords used in a particular disease or for a person and its frequency of occurrence made our analysis criteria even better.

Once again, the example of Cancer is considered. Using this disease, further screenshots of the results obtained are given. These are the results obtained through Cirrus Tool.Using the word frequency visualization tool (here Cirrus),the keyword density of certain most frequently occurring words can help in analysing the words mostly tweeted and there by helps know the trends through the size and frequency of its occurrence. Words are oriented either vertically or horizontally, and they are rendered in different colours just for ease of viewing and hence solves no further purpose regarding the position and placements as they do not indicate anything meaningful about the word. The size of the word, however, indicates the frequency with which it appears in the document. The larger the font size, the more frequent the word. The user can mouse over a word to see the precise number of times that it appears in the document. Here in figure 49 the word “cancer” is given the highest size which means that its occurrence in the tweets is the highest followed by other words which helps to know the keywords related to a particular disease, as an example see in figure 49 that words like awareness, breast, coffee etc. are also present which helps to relate these keywords to the disease specified in terms of its symptoms, type of disease and its treatments.

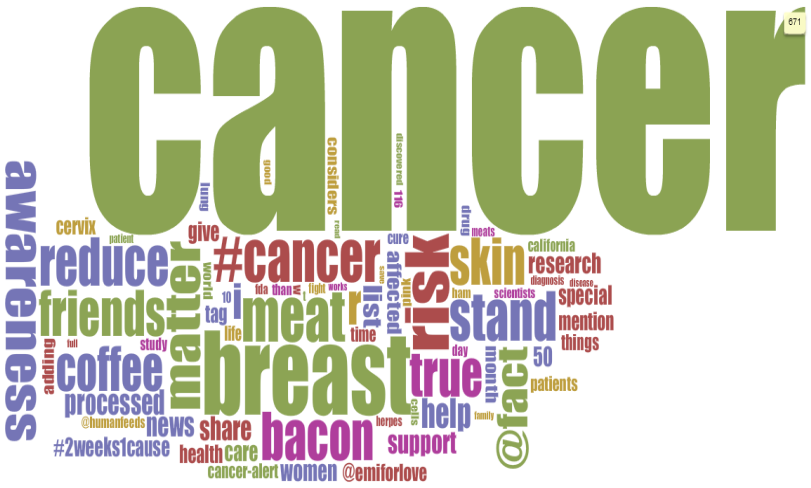

Figure 49: Word Cloud for extracted tweets

On clicking any particular word in the cloud (here cancer)we can get detailed analysis of that word in terms of "corpus reader", "word trend", "keywords in context" and "summary". In the summary part (which is bottom left in figure 50),the complete summary of all the most frequently occurring words is given in key value pair where the word name is specified along with its frequency in brackets. The word trends (which is top right in figure 50)shows the graphs of occurrence of the word in total number of tweets .Corpus reader and keywords in context shows the tweets where the word occurs.

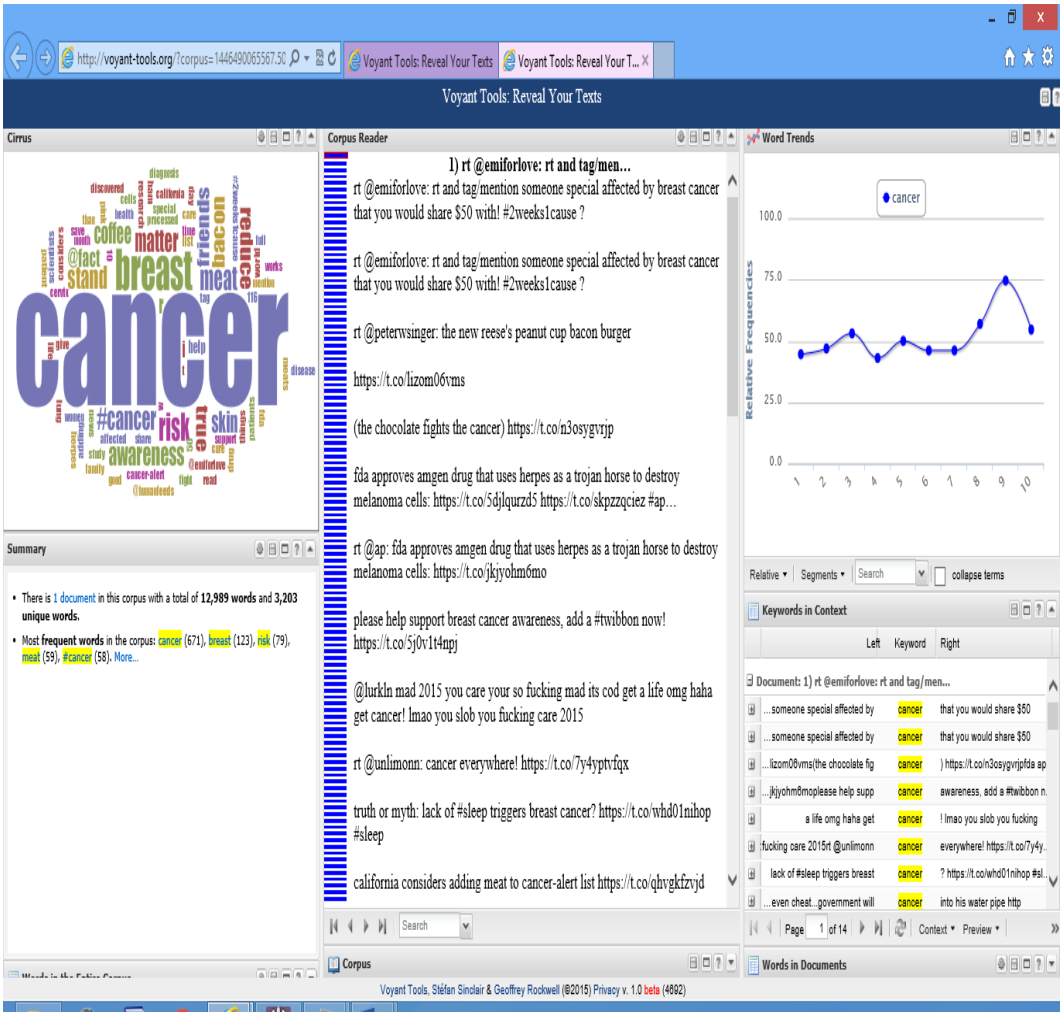

Figure 50: Cirrus Tool

4.3 Geo Mapping

Geo mapping in this project is done both dynamically and statically. In celebrity analysis we locate the locations in dynamic way whereas in health analysis we used static way

4.3.1Dynamic Mapping

The follower location which was extracted can be geolocated using maptive a tool.

Fig.51 shows the geolocation of followers located in different parts of the world.the label represents their position and the number on them represents the frequent of that particular region like 2 for ahemdabad

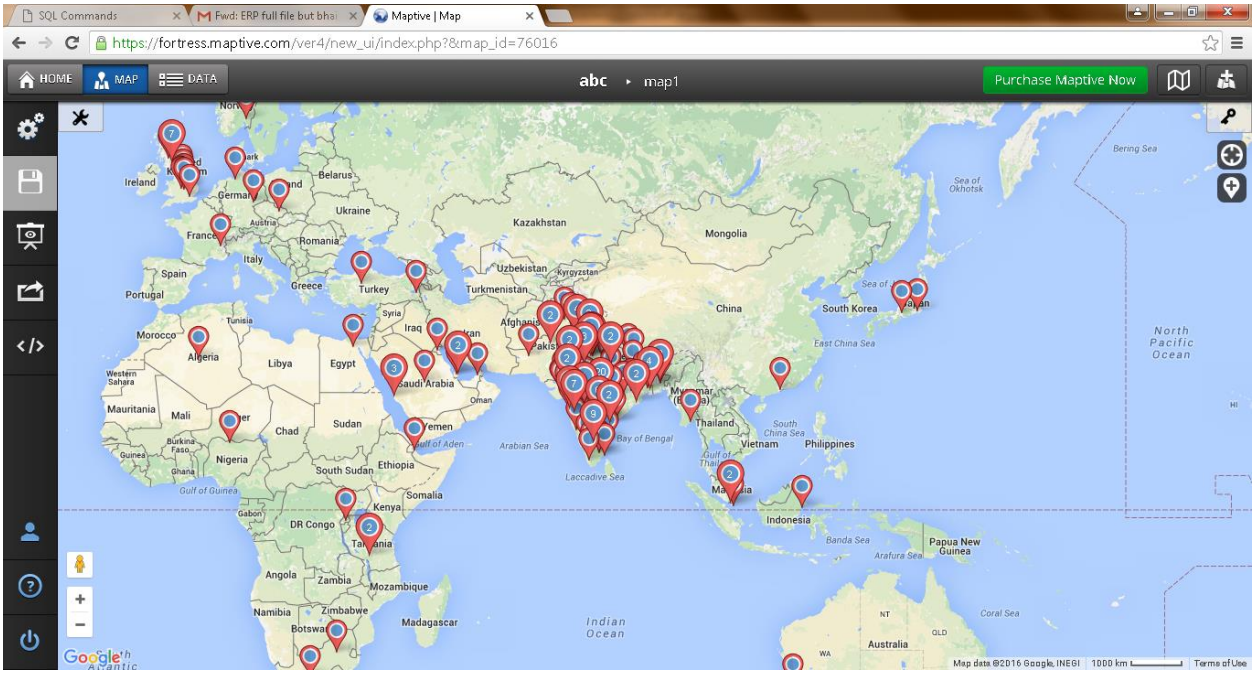

Figure 51:geolocation of followers

Fig. 52 represents the zoomed view of maps indicating that many of the followers are in india as compared to others.pakistan has less number.moreover terror countries like iran has no followers reflecting salman khan a better image. Fig. 53 represents that india is the country which is most popular among salman khan followers

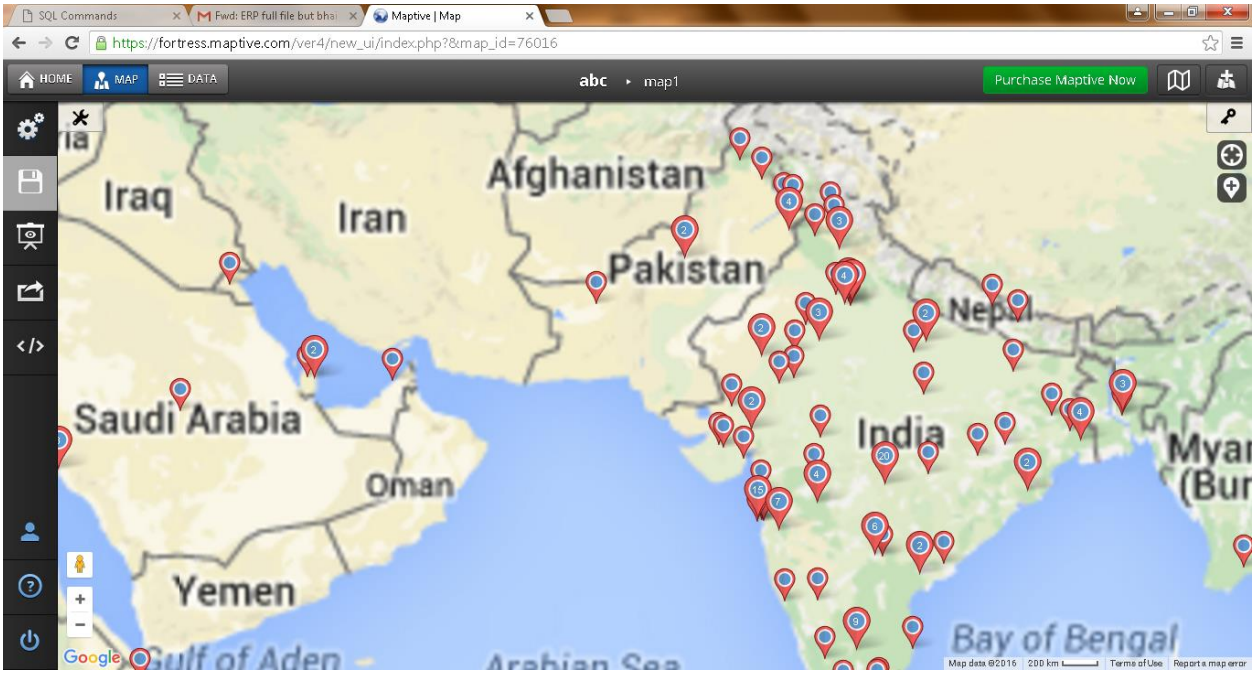

Figure 52:close view of geolocation

.

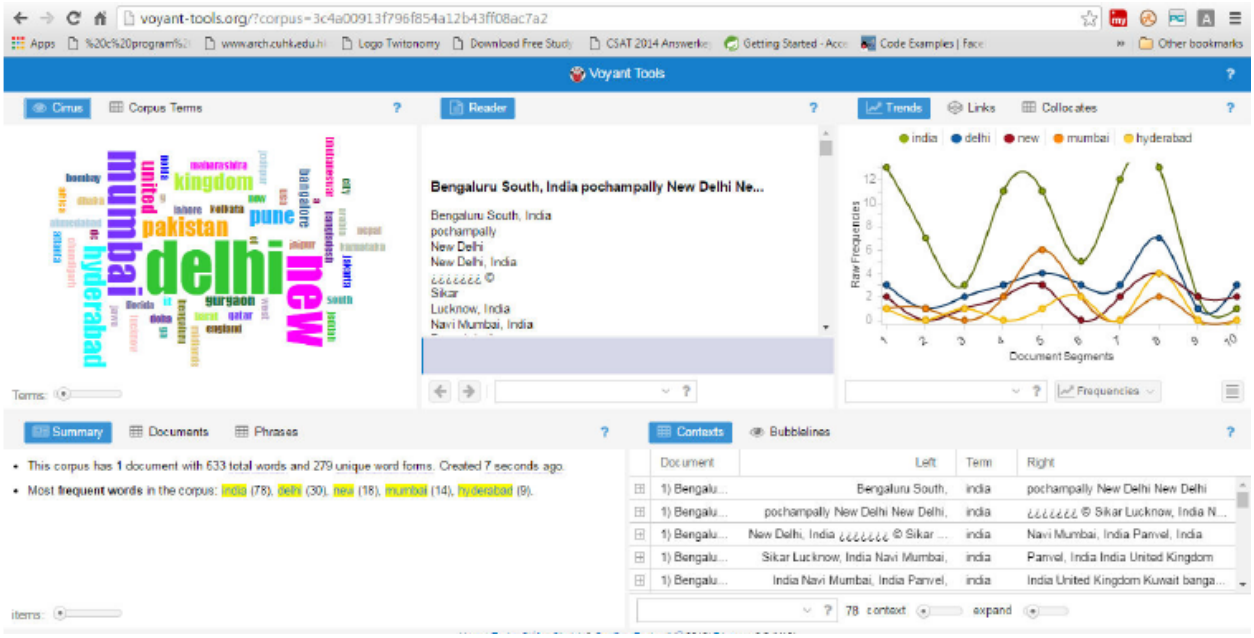

Figure 53:country wise analysis of followers

4.3.2Static Mapping

By using the database, a tweet frequency is calculated for each country for a particular disease. This calculated frequency is then used by JxBrowser and Google API service of maps marker for Geo-Mapping.

Figure 54 shows one such geo mapping. Balloon marks (marker position) represents the frequency of tweets for that particular area for example France: 1, marker has been placed at France and 1 represents its frequency, these marker has been placed on basis of their longitude and latitude. A tooltip is associated with one of the balloon representing the frequency of extracted tweets regarding a particular disease like 1 in case of France. Figure 55 shows the zoomed image for the same.

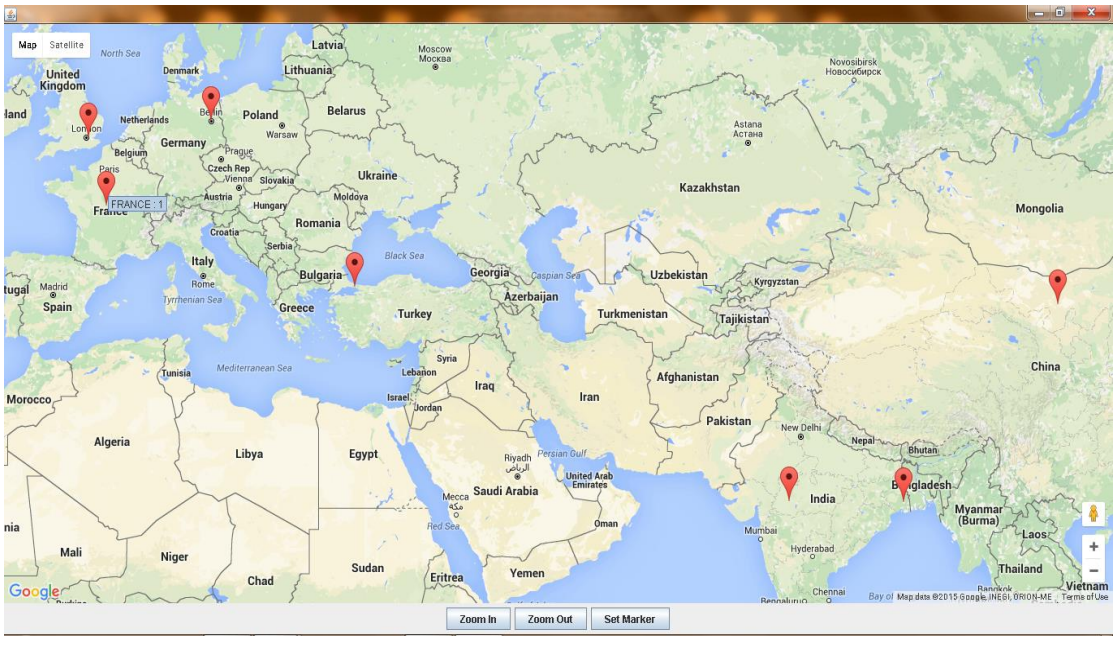

Figure 54: Geo Mapping

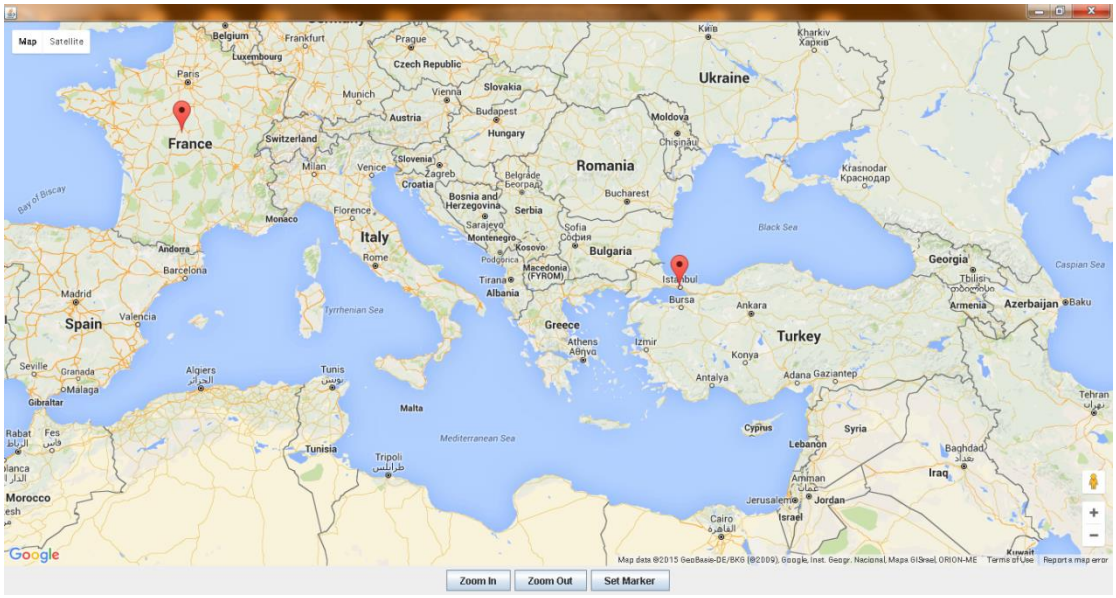

Figure 15: Zoomed in view

CHAPTER 5

CONCLUSION, SUMMARY AND FUTURE SCOPE

We designed and implemented a behavior analyzing system that uses twitter data to analyse behavior in real time. Twitter data is continuously downloaded using Twitter streaming API. Then the tweet texts are extracted along with the time stamps, and user locations and stored in a database, which is used for further analysis in the geographical and text terms. All the output data is visualized as an interactive map, bar charts, and word clouds. Our aim was to analyze the data present on social media. People often post messages about interesting news, their daily activities, thoughts and feelings. This may help to analyze their behavior and behavior of other peoples towards him/her. It makes easy to analyze positive and negative attitude from their tweets Twitter message, when used in combination, helps increase the accuracy of our prediction.
This project includes the extraction of tweets using Twitter API and then analyzing them in many possible ways. Apart from this, tools named RapidMiner, Cirrus, and maptive, plotly are used to take the analyzing process to another step.

We designed and implemented a novel disease surveillance system that uses twitter data to automatically track a particular disease activity in real time. We are also interested in using the social network information (e.g., friends and followers network) to predict the disease outbreak. We plan to gather additional online data (e.g., Facebook, blogs data, news feeds) for real-time disease surveillance. We set criteria to group certain symptoms and treatments in order to find the proximity of prevailing diseases and calamities

We will explore this project by studying more data mining tools such as twitonomy. And try to implement data mining algorithms such as Expectation Maximization (EM), Page Rank, AdaBoost and compare their accuracy with the other one.
Another possible research direction is to further integrate the modules into one Complete using Hadoop or R-programming. We may try to use Facebook as a data source to implement our analysis.
.

REFERENCES

1. Bo Pang and Lillian Lee. “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” 2004. Proceedings of ACL, pp. 271–278.
2. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs up? Sentiment Classification using Machine Learning Techniques.” 2002. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.
3. Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In COLING.
4. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, WWW ’10, pages 851
5. Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. Journal of Machine Learning Research (JMLR) 3.
6. Carneiro, H., and Mylonakis, E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49(10):1557–64.
7. Chew, C., and Eysenbach, G. 2010. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. PLoS ONE 5(11):e14118.
8. Culotta, A. 2010a. Detecting influenza epidemics by analyzing twitter messages. arXiv:1007.4748v1 [cs.IR].
9. Culotta, A. 2010b. Towards detecting influenza epidemics by analyzing twitter messages. In KDD Workshop on Social Media Analytics.
10. Eisenstein, J.; O’Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In Empirical Natural Language Processing Conference (EMNLP).
11. "Tweet Tracker: An Analysis Tool for Humanitarian and Disaster Relief". *The 5th International AAAI Conference on Weblogs and Social Media*.
12. Eysenbach, G. 2009. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res 11(1):e11.

13. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA, 2005.
14. Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter by Hassan Saif, Miriam Fernandez, Yulan He, Harith Alani Knowledge Media Institute, The Open University, UK
15. The Porter stemming algorithm: then and now. by Willett, P. (2006)a Program: electronic library and information systems, 40 (3). pp. 219-223.

16. J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. 2008. Detecting influenza epidemics using search engine query data. Nature, 457(7232):1012–1014.

17. A Survey of Algorithms for Keyword Search on Graph Data -By Haixun Wang and Charu C. Aggarwal
18. Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In IAPR 2nd Workshop on Cognitive Information Processing (CIP 2010).
19. Google flu trends. http://www.google.org/ flutrends, 2012.
20. H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on, 2011.
21. M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, ICWSM. The AAAI Press, 2011
22. Google chart tools. https://developers.google. com/chart/interactive/docs/gallery.
23. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. PLoS ONE, 6(5):e19467, 05 2011.