SENTIMENT ANALYSIS OF TWEET DATA

A Project Report submitted to the

Department of Computer Science and Engineering, Jahangirnagar University

in partial fulfillment of the requirements for the degree of

M.Sc. in Computer Science and Engineering

Submitted By:

Md. Delwar Hossain

Roll: CSE201702103

Batch: 14th

Session: 2017 - 2018

Supervised By:

Sanjit Kumar Saha
Assistant Professor
Department of Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JAHANGIRNAGAR UNIVERSITY

SAVAR, DHAKA – 1342
BANGLADESH

JULY 2018

DECLARATION

The research work entitled "**SENTIMENT ANALYSIS OF TWEET DATA**" has been carried out in the Department of Computer Science and Engineering, Jahangirnagar University is original and conforms the regulations of this University.

I understand the University's policy on plagiarism and declare that no part of this project has been copied from other sources or been previously submitted elsewhere for the award of any degree or diploma.

Submitted by-

_____

Md. Delwar Hossain
Student ID: CSE 201702103, PMSCS Batch: 14
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

APPROVAL OF ACCEPTANCE

A project report written by **Md. Delwar Hossain** (Student ID: CSE 201702103) entitled **"SENTIMENT ANALYSIS OF TWEET DATA"** is submitted to the PMSCS Program, Department of Computer Science and Engineering, Jahangirnagar University in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The project is done under the supervision of **Sanjit Kumar Saha,**Assistant Professor**,** Professor of Computer Science and Engineering Department, Jahangirnagar University, Savar, Dhaka.

We have examined this report and recommend its acceptance

....
Sanjit Kumar Saha
Supervisor and 1st Examiner
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

....
Md.Abul Kalam Azad
Professor & Coordinator
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Accepted by-

....
Md.Abul Kalam Azad
Professor & Coordinator
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessing makes me possible to complete this project successfully.

I feel grateful to and wish my profound my indebtedness to of **Sanjit Kumar Saha** Assistant Professor, Department of Computer Science and Engineering, Jahangirnagar University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of data mining influenced me to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to other faculty member and the staff of Computer Science and Engineering Department of Jahangirnagar University, for their kind helps to finish my project.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

This project is on "**SENTIMENT ANALYSIS OF TWEET DATA"** is a sentiment analysis project based on big data analytics. This project will help us to analyze sentiment from twitter text data. This will be able to show the sentimental state for a person or a specific topic or whatever I want. The aim of this project is to build up an application which will be able to connect non-technical people to the field of data mining. Using simple user interface, they will be able to create a report based on their search keyword. This application is very simple and easy to use. So, it will be comfortable for everyone to use it. To develop this project, need to know R language and platform and its library. To create the user interface using shiny, which is a package including PHP and HTML to create a web application. By giving input inside the shiny app and in the background, R will do the data collection, analysis and output generation. Then shiny will print the output in the web application. Its main task is running the application on the browser, which is taking input and providing us output based on input.

TABLE OF CONTENTS

CONTENS   **PAGE**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Twitter, a popular social network site, is one such source which contributes huge data which possess values beyond social and commercial interests. Twitter users can express or share their opinions, feelings or information regarding events, products, health or anything in their 140-character restricted short messages termed tweets. It is a great source of data. Nowadays, everyday 500 million tweets are posted in twitter. So, it has become very important to analyze them to find track of data flowing throw twitter. As a part of data analyzing data or data mining analyzing the sentiment of twitter data is also important to know the sentiment of the user. Analyzed data can be used for research, business market study, product review etc.

1.1 OBJECTIVE

Research is an organized investment of a problem in which there is an attempt to gain solution to a problem. To get right solution of a right problem, clearly defined objectives are very important. Clearly defined objectives enlighten the way in which the researcher has to proceed. Research objectives are usually expressed in lay terms and are directed as much to the client as to the researcher. Research objectives may be linked with a hypothesis or used as a statement of purpose in a study that does not have a hypothesis. A researcher objective is clear, concise, declarative statement, which provides direction to investigate the variables. Generally, research objective focus on the way to measure the variable, such as identify or describe them. Sometime objectives are directed towards identifying the relationship difference between two variables. Research objective outline the specific goals the study plans to achieve when completed.

My research objectives are given below:

- To study Big Data & Data Analytics.

- To study sentiment analysis.

- To develop and seek Knowledge on how to deal with tweet data.

- To develop Computational tool for sentiment analysis.

- To apply the knowledge of sentiment analysis on the real-life example.

- To build an app for non-technical users.

1.2 MOTIVATION

The explosive growth of the textual information on the web in the past few decades has brought radical change in human life. In the web, people share their opinions and sentiments. In many forms about services the services they are aware of. This creates a large collection of opinions and views in the form of texts, which needs to be analyzed to know the efficacy of the service. Opinions are a usually subjective expression that describes person's sentiment, feelings towards the object or service.

Now a day's twitter is a very popular social website were from all over the world people connect with each other and communicate and express their thought through texts. By analyzing their tweets, I can decide whatever their mental situation is going on. Even what is the recent "hot topic". So, I was motivated to analyze tweet data from this thinking. Before starting on something motivation is very important. Because without proper motivation it's not possible to carry on with it. In some cases, only motivation can complete a huge part of an analysis or task.

Motivation for my projects are:

- Recent revolution of data analysis and big data.
- My personal interest on data mining and sentiment analysis.
- Opportunity to learn new things.

Jahangirnagar University

1.3 EXPECTED OUTCOME

I choose "Sentiment analysis of tweet data" as my project because I wanted to build a data mining app which will be able to create interaction between tweet data and data mining. I wanted to build an app which will be able to collect twitter data for example tweets and analyze them. After the analysis the application will provide us the statistical overview of sentimental situation of targeted people.

So finally, I decided to build a data mining application using R for sentiment analysis. Where I will be able to search tweets using keyword, for a particular topic, of a particular account or random data. User will also be able to see trending hot topics or keywords on their search topic. I will also build a web application for non-technical users. So that they will be able to analyze data themselves from twitter. Outputs will be viewed as histogram plot, bar diagram and word cloud.

1.4 REPORT LAYOUT

There are five chapters in this research paper. They are: Introduction, Literature Review, Proposed Research, Results and Discussion, Conclusion and Future.

Chapter one: Introduction; Objective, Motivation, Expected Outcome, Report layout.

Chapter two: Literature Review; Sentiment Analysis, Related works, challenges.

Chapter three: Proposed Research; Research Methodology, Data Collection, Data Processing, Flow Model, Experimental layout.

Chapter four: Results and Discussion; Experimental Result, Discussion.

Chapter five: Conclusion and Future; conclusion, Future Scope.

CHAPTER 2

LITERATURE REVIEW

Sentiment analysis is study of data and finding out the sentimental state. Data can be text or a audio or a video file. Sentiment can be measured by applying different method for different type of data. For example, text can be analyzed by the keywords on it, audio files can be analyzed using tone of the voice and video or image file can be analyzed by tracking the state of the face. Recently sentiment analysis has become very popular because of the large data flow over the internet. Researchers are trying to analyze different type of data for different purpose. But the main challenge for sentiment analysis is detecting actual state of emotion. Because some cases analyzed data don't result what they actually mean. Because some positive sentences sometimes mean as negative and some negatives mean as positive. Researchers are now trying to solve this kind of issues to get much more accurate result of sentiment analysis.

2.1 SENTIMENT ANALYSIS

Sentiment means feelings, it also be attitudes, emotions or opinions. And sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Using natural language processing, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit. Sometimes it referred to as opinion mining.

The purpose of sentiment analysis is to determine the attitude of a communicator through the contextual polarity of their speaking or writing. Their attitude may be reflected in

their own judgment, emotional state of the subject, or the state of any emotional communication they are using to affect a reader or listener.

Sentiment analysis is a fast-growing subject in the technical communication field. With the increase in social media, online retail, and personal blogs and publications knowing where public sentiment is leaning has translated into a rapid evolution in sentiment analysis that can become a valuable skill.

When I perform sentiment analysis on some content, the main searching point is the opinions in content and I picking the sentiment based on those opinions. An opinion is an expression that consists of two key components. One of them is a target or topic and another one is a sentiment on the topic.

Consider a sentence, "I love this company ", here "this company" is the topic and the sentiment that is expressed by the verb "love" is positive.

Sentiment analysis is just not a feature in a social analytics tool – it's a field of study. This field is still being studied, albeit not at great lengths due to the intricacy of this analysis, in the same way that some aspects of linguistics are still up to debate or not fully understood.

Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches.

Knowledge-based techniques classify text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Some knowledge bases not only list obvious affect words, but also assign arbitrary words a probable "affinity" to particular emotions.

Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation — Point

wise Mutual Information. More sophisticated methods try to detect the holder of a sentiment (the person who maintains that affective state) and the target (the entity about which the affect is felt). To mine the opinion in context and get the feature which has been opinionated, the grammatical relationships of words are used. Grammatical dependency relations are obtained by deep parsing of the text.

Hybrid approaches leverage on both machine learning and elements from knowledge representation such as ontology and semantic networks in order to detect semantics that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

Open source software tools deploy machine learning, statistics, and natural language processing techniques to automate sentiment analysis on large collections of texts, including web pages, online news, internet discussion groups, online reviews, web blogs, and social media. Knowledge-based systems, on the other hand, make use of publicly available resources, to extract the semantic and affective information associated with natural language concepts. Sentiment analysis can also be performed on visual content, i.e., images and videos. One of the first approach in this direction is SENTIBANK utilizing an adjective noun pair representation of visual content. [1]

The applications for sentiment analysis are endless. More and more I am seeing it used in social media monitoring and VOC to track customer reviews, survey responses, competitors, etc. However, it is also practical for use in business analytics and situations in which text needs to be analyzed.

2.2 RELATED WORKS

Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment (and other features including named entities, topics, themes, etc.) in seconds, compared to the hours it would take a team of people to

manually complete. Because it is so efficient (and accurate – Semantria has 80% accuracy for English content) many businesses are adopting text and sentiment analysis and incorporating it into their processes. [2]

People all around the world nowadays working on different topics of sentiment analysis. As the amount of data increasing day by day data mining and sentiment analysis became more popular to people. Different companies need sentiment analysis for collecting customer information. Because this is the easiest way to find their customers and target them for their business.

Even now social media also need sentiment analysis. Using sentiment analysis, they determine marketing strategy, improve campaign success, improve product messaging and improve customer service.

Some sentiment analysis related works are:

1. Opinion mining and sentiment analysis. By Bo Pang and Lillian Lee, Foundations and Trends in Information Retrieval, Vol. 2, No 1-2 (2008) 1–135.
2. Twitter Sentiment Analysis: The Good the Bad and the OMG! ByEfthymios Kouloumpis, Theresa Wilson, Johanna Moore, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
3. Sentiment Analysis of Short Informal Texts, Svetlana Kiritchenko, Xiaodan Zhu and Saif M. Mohammad, Journal of Artificial Intelligence Research 50 (2014) 723–762.
4. Sentiment Analysis: Capturing Favorability Using Natural Language Processing, Tetsuya Nasukawa and Jeonghee Yi.

5. Sentiment Analysis of Political Tweets: Towards an Accurate Classifier, Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi and Mark Hughes.

6. Stock trend prediction using news sentiment analysis, Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao.

7. Decision Making Using Sentiment Analysis from Twitter, M.Vasuki, J.Arthi, K.Kayalvizh.

All those are the recent research works on sentiment analysis. They focused on different topics like customer review, decision making, NLP, twitter data, openion mining etc.

2.3 CHALLENGES

As I know that sentiment analysis is the analysis of image, text, audio, video sentiment recognition. Sentiment analysis and evaluation process are facing several challenges. These challenges become obstacles in analyzing the accurate meaning of sentiments and detecting the suitable sentiment polarity. To identify and extract subjective information from text the sentiment analysis is the practice of applying natural language processing and text analysis techniques [3].

The degree of accuracy issue is hard to answer, said Bing Liu, a University of Chicago computer science professor specializing in data mining. It depends on what are measuring, the level of analyzing text, and the number of data sets across domains and the voice sound quality of videos, among other variables. Still, he thinks that progress is being made in this regard [2].

In sentiment analysis it is more challenging to detect a more in depth sentiment/emotion. Positive and negative is a very simple analysis but the challenging one is to extract

emotions like how much hate there is inside the opinion, how much happiness, how much sadness, etc.

Emotion detection is really a difficult task because sometimes it happens that someone tell something that seems positive but in real it's not positive the sense was negative. So sometimes it is difficult to understand meaning of a sentence cause the emotions are too much complex.

Mainly sentiment analysis tries to detect the mental situation of a person. But sometimes it become tough to tell what the person meant. If I consider audio sentiment analysis, then noise or voice tune difference can create major error in output. Same for text analysis, because some texts word wise meaning is totally different from its actual meaning. That's why sentiment analysis is facing major challenges nowadays.

CHAPTER 3

PROPOSED RESEARCH

In this project I try to analyze twitter data to get sentimental state from it. To do that at first, I collect data from twitter. I create a twitter app to get the data. Then connect it with my application. I write a script to analyze collected data. So, when run the program it will automatically collect the data, analyze it and give the output.

3.1 RESEARCH METHODOLOGY

As I select sentiment analysis of tweet data as my project. So, I collect data from twitter. Everyday twitter generates tons of or terabytes of data by its users. Where most of them are unstructured data. So before starting my research work and implementation I need to consider this as my challenge for sentiment analysis.

I find different platforms for data mining as orange, Weka, Rattle GUI, Apache Mahout, R, Hadoop, UIMA, SenticNet API, Natural Language Toolkit etc. Then I selected R as my data mining tool and platform.

I select R as my data mining tool and platform because of its friendly user interface and strong library for data processing, data mining and output visualization. So, at first, I learnt how to work with R. Then learning R programming language which is mainly used for data analysis and it is a high-level programing language. I learned R programming language from https://www.datacamp.com. It is very helpful website because its shows how to learn R in step by step process.
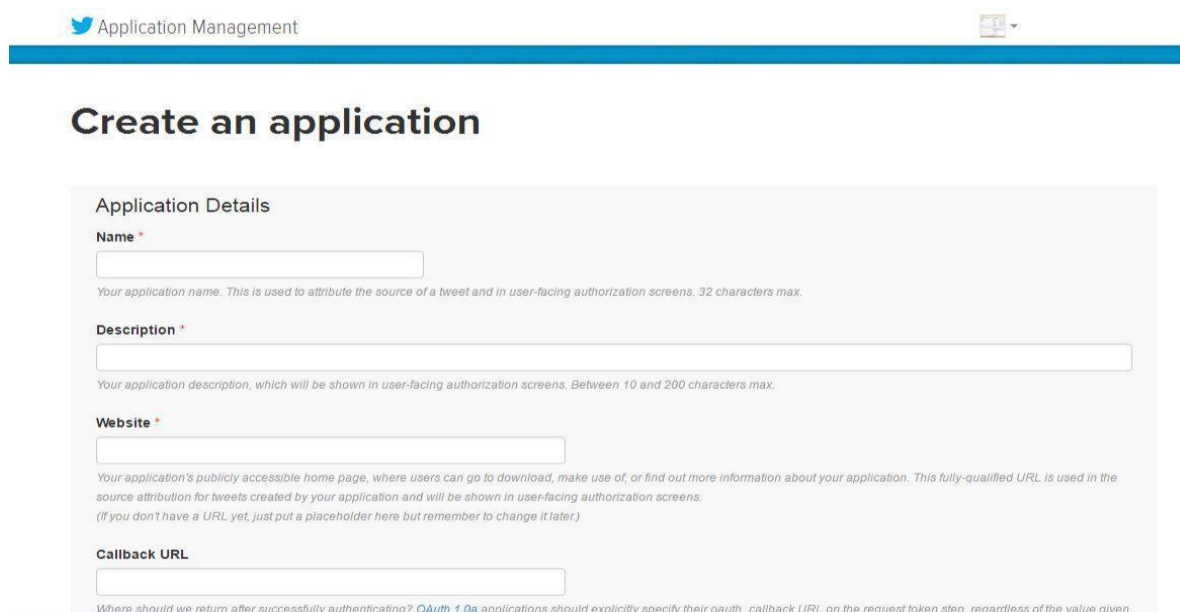
After that I create a twitter app for data collection and using R language. This application can connect to twitter using internet and collect data and analyze them. Collected data are stored in a CSV file.

3.2 DATA COLLECTION

To collect data, I needed a twitter application, which can give me access to the twitter and application will be able to find expected tweets. To create a twitter application, I went to the twitter application management website. Then I create an application named Sentiment Analysis BDA.

3.2.1 CREATING TWITTER APP

To create the application, I need to fill up a form providing name of the application, what is the purpose of the application and other necessary data.

Figure 3.1 creating a twitter application

After providing all information my application is created and now I can access it and view detail information about it.

Jahangirnagar University

Figure 3.2 application created in twitter

In this application home page, they provided us some important information called consumer key, consumer, Consumer Secret, access token, access token secret about accessing the app to collect tweet. Without accessing that information no one will be able to connect this app with my application.

Jahangirnagar University

Figure 3.3 application home page

3.2.2 R SETUP

When creating twitter application is done now it is ready to build R application to collect data and store them. To create R application first download and instal the R application to my computer.

Figure 3.4 R download

Jahangirnagar University

After downloading and installing the R app I will get a user interface shown below:
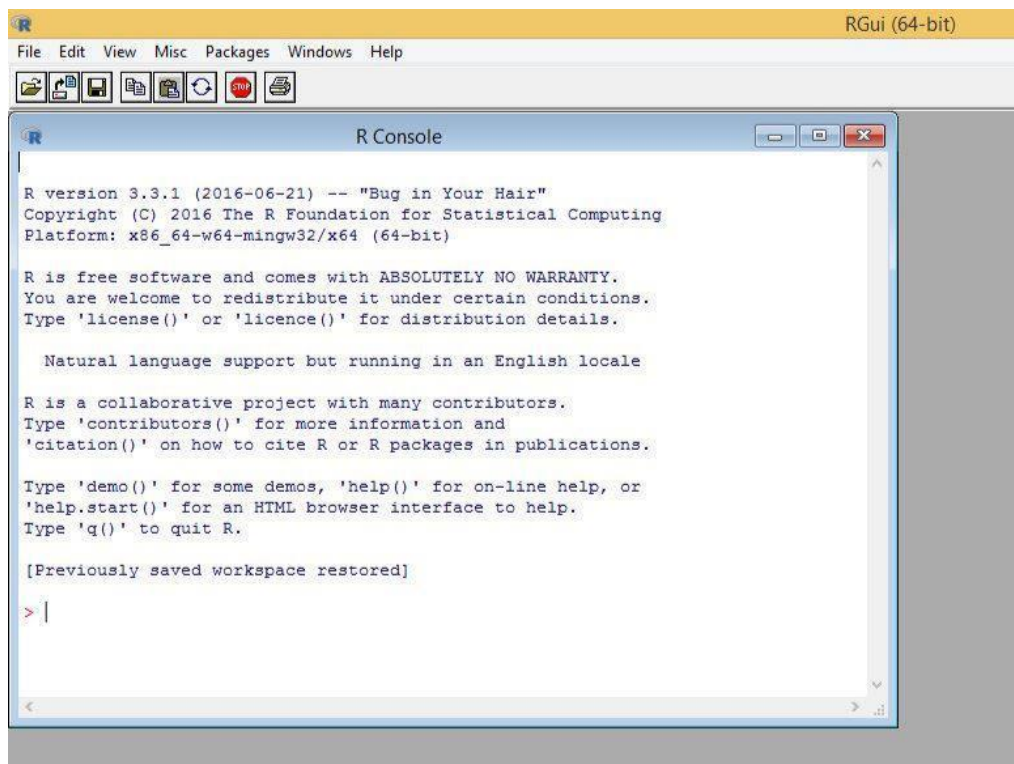
Figure 3.5 R user interface

But the user interface doesn't look that much comfortable to use and there is no visualization option. Then I consider to install another application named RStudio.

RStudio gives R programmers a very strong and helpful user interface. They will also get a visualization platform where they will be able to visualize their statistical result or output.

Jahangirnagar University
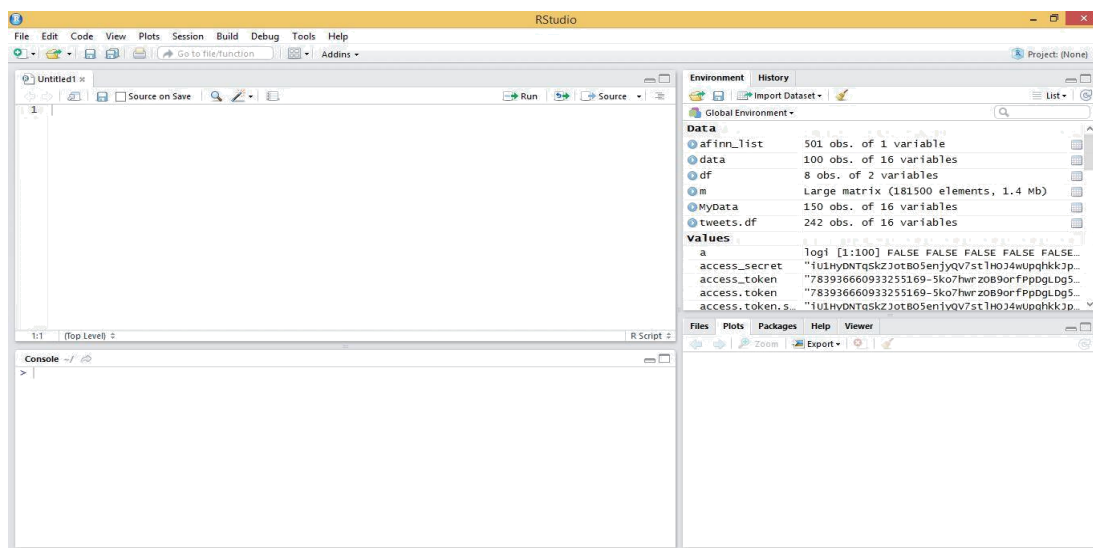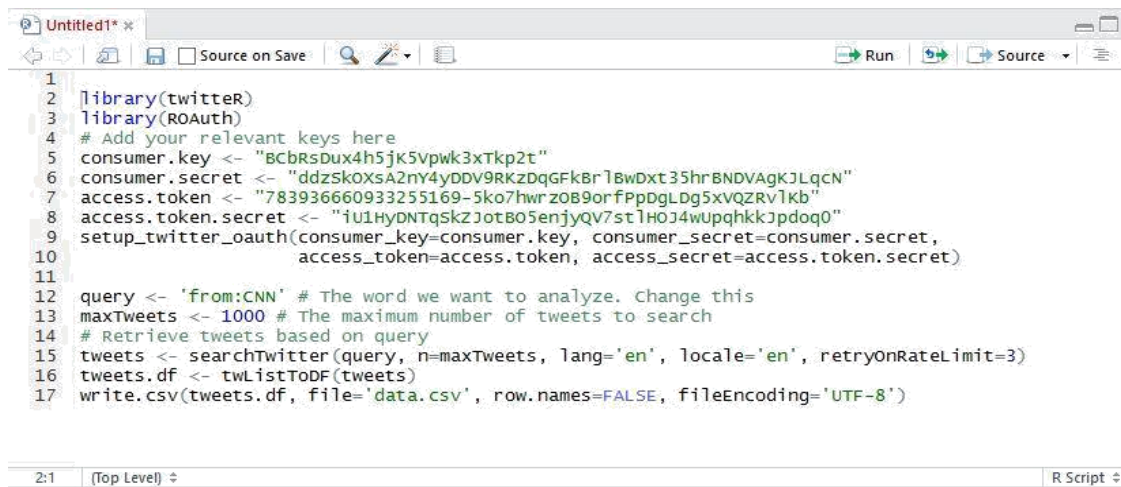
Figure 3.6 RStudio user interface

After installing necessary components now, it is ready to write the program to collect tweets and store them.

3.2.3 COLLECTING TWITTER DATA

To collect data from twitter I have to write a script which will connect the application and collect tweets from twitter. Now the script for data collection and connection is given below:

Figure 3.7 data collection and store

I used two libraries library(twitteR), library(ROAuth) for data collection and linking the application with twitter.This script will collect data from twitter and store them in a CSV file. But as I need to analyze them and show user the output so I later only worked with data file collected from twitter and analyzed them.

Tweets will be collected based on last upload. User will be able to collect as much as tweets he wants. Tweet collection will be based on search keyword, number of tweets and existence of that tweet keyword in the twitter.

3.3 DATA PROCESSING

When all the data are collected, my main task then starts. This is "Sentiment analysis of tweet data". But for processing the data I need to work in two steps. They are:

   i.   Data pre-processing
  ii.   Data analyze

After completing those steps, I will be able to get my expected output of my project.

Jahangirnagar University

3.3.1 DATA PRE-PROCESSING

In this step I will cut off all the unexpected and unwanted data generated by twitter as location, retweets number, link etc.

To cut off unexpected data I need the following script:

```
1
2  sentence <- gsub("[^[:alnum:]///' ]", "", sentence)
3  sentence <- gsub('[[:punct:]]', '', sentence)
4  sentence <- gsub('[[:cntrl:]]', '', sentence)
5  sentence <- gsub('\\d+', '', sentence)
6  sentence <- tolower(sentence)
7  word.list <- str_split(sentence, '\\s+')
8  words <- unlist(word.list)
```

Figure 3.8 shorting data

Those codes will short unwanted columns and only store the tweets. Data pre-processing is important because without pre-processing I won't get better output and analyzing process will take more time.

3.3.2 DATA ANALYZE

When tweets are pre-processed data are ready to analyze. There is several methods to analyze data for sentiment analysis. They are Sentiment Classification Using Lexical Contextual Sentence Structure, Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis, Interdependent Latent Dirichlet Allocation, A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating, Opinion Digger, Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach.[5]

In my project I used combining Lexicon and Learning based approaches for concept-level sentiment analysis. Here I first divided my tweet sentences into words. So that I are able to compare with my stored positive and negative words.

I used positive and negative words as my word library for positive and negative sentiment. I stored them as different .txt file and stored them in the same directory where the r and CSV files are stored. Demo of positive and negative words are given below:

Figure 3.9 positive and negative word library

Now I need to write an R script which will read my pre-processed tweet data and divide them into words. Then it will match those words with my previously stored positive and negative words which I are considering as my library for my sentiment analysis.

Here the script will also score my tweets according to its positive or negative values. If the positive value is greater than negative value the tweet is a positive tweet.

Jahangirnagar University

The scoring system works on based on following method:
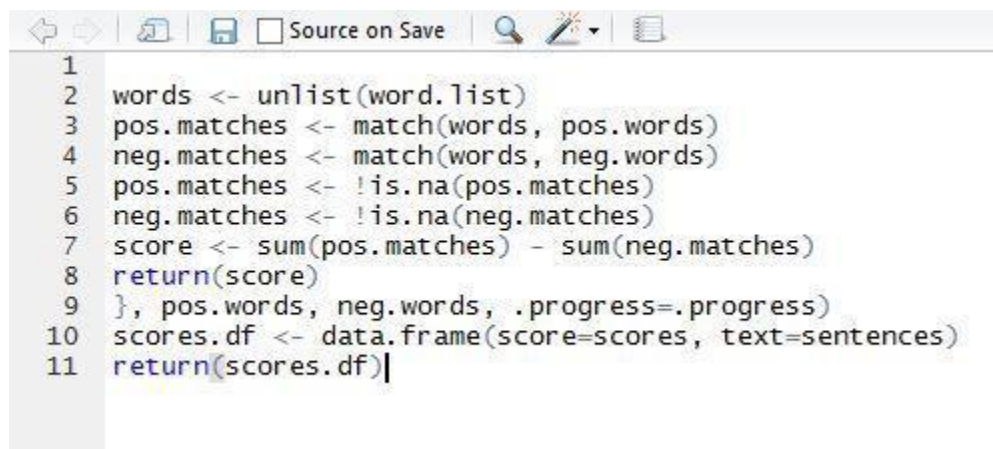
```
positive = 0;
negative = 0;
neutral = 0;

if(word == positive_word)
{
        positive++;
}
else if(word == negative_word)
{
        negative++;
}
else
{
        neutral++;
}

tweet_value = positive - negative;
```

Figure 3.10 sentiment scoring method

Now I will apply this same method in R to analyze the tweet data which I collected from twitter before and processed. The R program for analyzing data is given below:

```
1
2  words <- unlist(word.list)
3  pos.matches <- match(words, pos.words)
4  neg.matches <- match(words, neg.words)
5  pos.matches <- !is.na(pos.matches)
6  neg.matches <- !is.na(neg.matches)
7  score <- sum(pos.matches) - sum(neg.matches)
8  return(score)
9  }, pos.words, neg.words, .progress=.progress)
10 scores.df <- data.frame(score=scores, text=sentences)
11 return(scores.df)
```

Figure 3.11 Tweet sentiment scoring

19

After calculating the twitter value, I will get the twitter value as positive or negative. If the tweet is not positive or negative then it will be counted as neutral tweet.

3.4 FLOW MODEL

In my project I divided my data flow is different parts according to how data are passing through each other. First a twitter app is connected with a twitter account, then an R script connect it with R, so now data collection is ready. When data collection is complete then R script analyze it and gives some statistical report as bar diagram, word cloud as output. Than all the outputs are sent to the browser and browser visualize them for user.
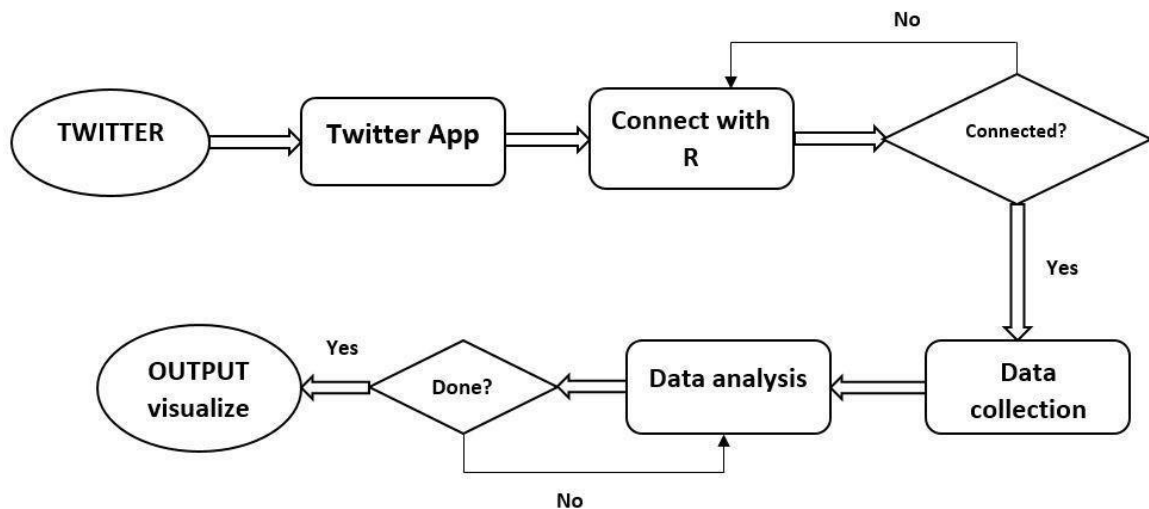
Figure 3.12 Data flow

According to the flow model system will recheck for data connection before collecting data. If it is not connected then it will try total 3 times then connection will fail. Data flow will be direct as it is given is the figure.

Jahangirnagar University

3.5 EXPERIMENTAL LAYOUT

According to my plan I wanted to build a user interface for general users who wants to analyze the tweet data. So first of all I need a search box, which will give us access for searching tweets. Then I need an output visualization part for viewing the results. I will have multiple outputs for the search results.

So I need to plot them in different places so that user can easily understand the result generated by the program. As well as outputs must change depending on different search keyword. So user don't need to reopen the app again and again.

I are going to use shiny. So as shiny default there is three parts in a web page.

1. Title panel,
2. Side bar panel,
3. Main panel.

So considering all the conditions I designed a layout for my project. The layout is given below:
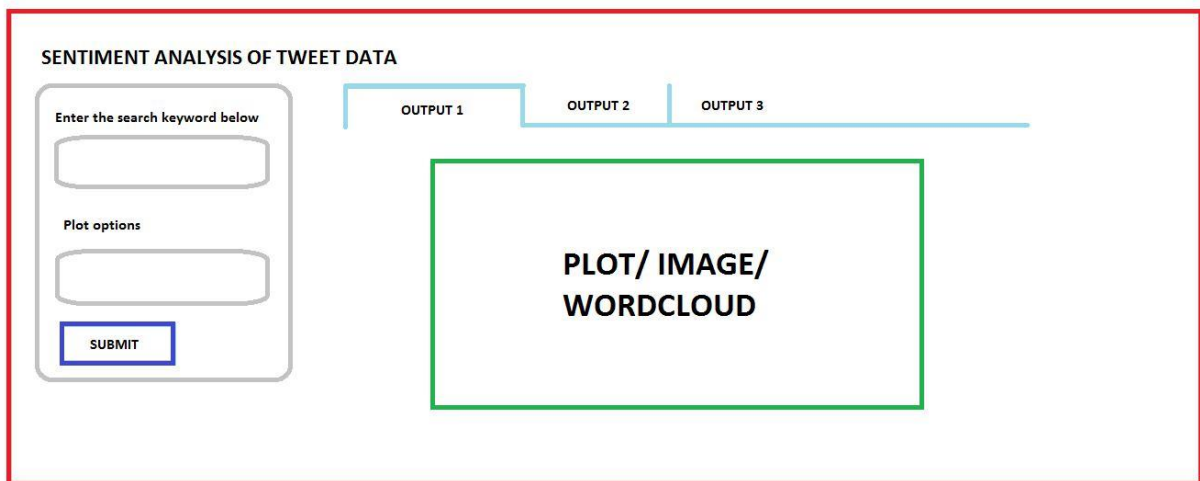
Figure 3.13 Experimental Layout

CHAPTER 4
RESULTS AND DISCUSSION

To use an application creating a user interface is the best option. So, I create a user interface for my application so that anyone can use them easily. I use two types of algorithms as Naive Bayes and Laxicon, on the other hand I got four kinds of data as output. All the outputs are visualized in my application via web browser.

4.1 EXPERIMENTAL RESULT

I created my project in two different ways for both who knows about R programming language and who don't know anything about it. Because there is always some people who are interested about the internal structure and play with it and some other people who wants to work using a simple user interface.

So first I created a program that can connect my program with R then it will collect data and store them in the working directory. After storing the data my program will retrieve the store data and short them and it is ready for analysis. In analysis part it will happen in 4 different steps.

1. Emotion,
2. Polarity,
3. WordCloud &
4. Scoring.

Those analysis will show us the complete state of the sentiment for my collected tweets. I stored every kind of data like CSV, image etc. Because by storing them if I want to do some more analysis, it will be a lot easier. For example, if I want to compare all the news of CNN for the month January and February then I just need to collect the data and store them after processing in different location and compare them using the WordCloud and bar charts.

4.1.1 EMOTION CLASS

An emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response. Emotions, can be considered as feelings, include experiences such as love, hate, anger, trust, joy, panic, fear, and grief. Emotions are related to, but different from, mood.

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".[6]

There are 6 emotion categories that are widely used to describe humans' basic emotions, based on facial expression [7]: anger, disgust, fear, happiness, sadness and surprise. These are mainly associated with negative sentiment, with "Surprise" being the most ambiguous, as it can be associated with either positive or negative feelings. Interestingly, the number of basic human emotions has been recently "reduced", or rather re-categorized, to just 4; happiness, sadness, fear/surprise, and anger/disgust.[8]

So if I want to do sentiment analysis at first I need to know and understand what emotion is and how it can be described or expressed. Without this analyzing human sentiment through tweets will be difficult to do.

For my twitter sentiment analysis project I divided tweet sentiments in 6 important classes. They are joy, sadness, surprise, anger, fear and unknown. To do this I used Naive Bayes algorithm (Bag of Words) so that I can find out the individual scores for all those emotion classes. Between all of them unknown class was calculated by the unknown tweets number.

Jahangirnagar University

To get the emotion class I need two libraries. They are require(plyr) and require(stringr). Where plyr is a tool for Tools for splitting, applying and combining Data and stringr is a tool for simple, consistent wrappers for common string operations.

To analyze the emotion of the tweets at first I need to consider my preprocessed data for applying Naive Bayes algorithm. This algorithm can analyze data for two state polarity and emotion. So here I will run the emotion program, so that I can find out emotional value of the tweets.

I wrote a simple R script to apply Naive Bayes algorithm for emotion analysis. The script is given below:

```
#this function performs emotion sentiment analysis using Naive Bayes Classification
#input : text(tweets)
#output : emotionclass

emotionSentimentAnalysis <- function (inText) {

  emotionclass <-  classify_emotion(inText, algorithm="bayes", prior=1.0)

  #extract emotion with the best possible fit
  emotion <-  emotionclass[,7]

  #setting emaotion having NA to unkonw
  emotion[is.na(emotion)] = "unknown"

  return (emotion)
}
```

Figure 4.1 Emotion class sudo code

When I run the program, it automatically analyze the tweets and apply Naïve Bayes algorithm to analyze emotion. This algorithm mainly classify emotions in 5 basic classes. They are anger, fear, joy, sadness and surprise. After that rest of the tweets are counted as Unknown class.

Jahangirnagar University

After running the program, I get the output as a bar diagram. I set different color for different bars for better visualization. The output bar plot is given below:
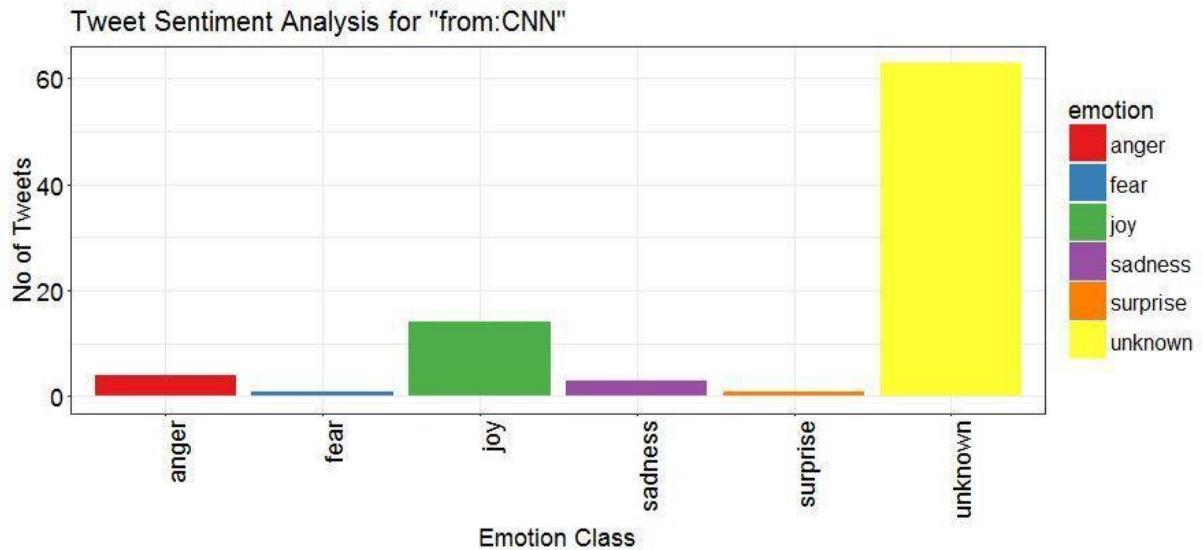
Figure 4.2 Emotion class bar diagram

In following figure, I got output for the channel CNN. But most of the tweets were counted as unknown because Naïve Bayes algorithm could not find out the class for those tweets. I also can see which emotion class having which value or how many tweets they have.

4.1.2 POLARITY CLASS

Polarity is the state of emotion which defines the positivity or negativity of text. Polarity, also known as orientation is he emotion expressed in the sentence. A text can be divided into three forms, positive, negative and neutral. Depending on the comparison with the positive and negative words score of a tweet can be easily counted.

Jahangirnagar University

As like as emotion class I applied Naïve Bayes algorithm in here too. As I know the basic task of sentiment analysis is figure out the sentiment state for a text depending on the words in it.

To find out the priority at first, I need to divide the text into single words so that comparison or the calculation become easier. Then Naïve Bayes algorithm search for words that can express sentiment in a tweet.

As I know Naïve Bayes algorithm works with bag of words. So first of all, I need to find out all the positive or negative words.

For example, a list of positive and negative words is given below:

Positive words	Abound, abounds, acclamation, accolade, accolades, accommodative, accommodative, accomplish, accomplished, accomplishment, accomplishments, accurate, accurately, achievable, achievement, achievements, achievable
Negative words	2-faced, 2-faces, abnormal, abolish, abominable, abominably, abominate, abomination, abort, aborted, aborts, abrade, abrasive, abrupt, abruptly, abscond, absence, absent-minded, absentee, absurd, absurdity, absurdly, absurdness, abuse

Those bags of words are used to find out if the tweet is positive or negative. For each positive or negative word every tweet will get 1 score. Here,

Tweet score = tweets positive value – tweets negative value

For example, I have a tweet "that was a pretty good party."

For this I will get the score: positive = pretty (1) + good (1) = 2;

Negative = 0

So the tweet score for this tweet will be

Tweet score = 2 – 0 = 2

All the tweets having positive tweet score will be positive tweets and all the tweets having negative tweet score will be negative tweets. If any tweet contains tweet value 0, it will be under neutral tweet. Counting all of them I will get a proper chart of positive, negative and neutral tweets.

To calculate the polarity of the text I write an R script which can calculate the tweet score using Naïve Bayes algorithm. The R script is given below:

```
#This function Clasifiys polarity of the of the tweets bases on naive bayes algorithm
#Input : text(tweets)
#Output : polarityclass

polaritySentimentAnalysis <- function (inText) {
  #polarity clasification
  polarityclass <- classify_polarity(inText,algorithm = "bayes")

  #polarity best fit for plottin
  polarity <- polarityclass[,4]

  #setiing polarity having NA to unkown
  polarity[is.na(polarity)] <-"unknown"

  return(polarity)

}
```

Figure 4.3 Polarity class sudo code

When I run this program, it will find out all the positive and negative words and start analyzing them to differentiate them. So, they will get their tweet score. Then according to their tweet score they create a histogram plot of positive, negative and neutral tweets. After running the polarity class program, I will get the following diagram:

27
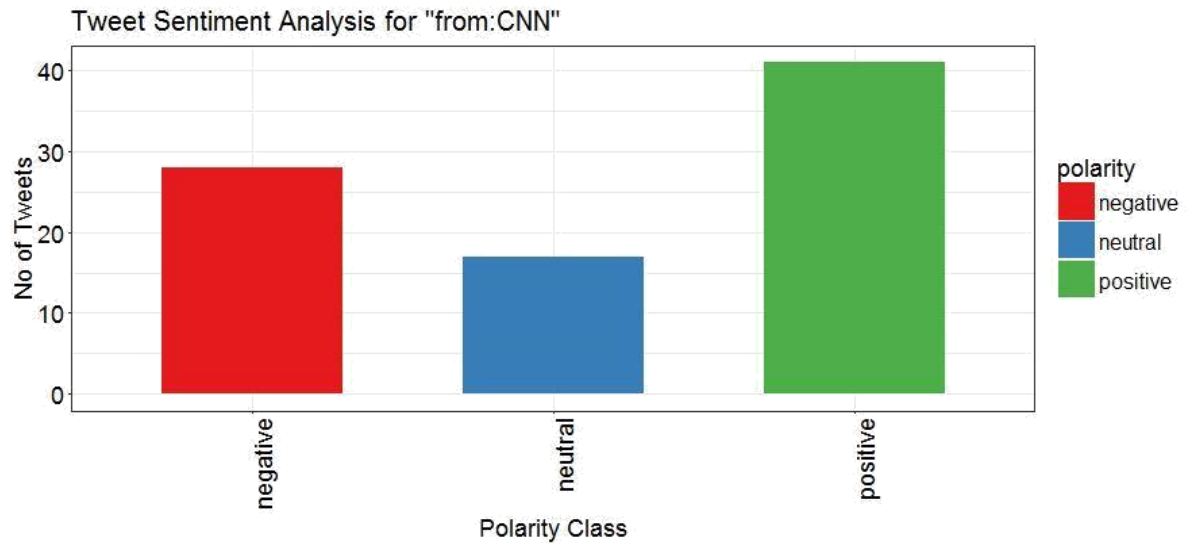
Figure 4.4 Polarity class bar diagram

4.1.3 WORD CLOUD

If I want to define word cloud then I can say that an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document(s). This type of visualization can assist evaluators with exploratory textual analysis by identifying words

28

that frequently appear in a set of interviews, documents, or other text. It can also be used for communicating the most salient points or themes in the reporting stage.[9]

So I also analyzed word cloud to find out hot topic or what about people are talking most now a days. Word cloud is very important to understand people's sentiment and what about they are sad or happy. Without considering word cloud it will be tough to find out key points of sentiment analysis bar chart or scoring.

Though there is a lot of tools to create a word cloud. Where I just need to upload my document file or the website link and they provide us the word cloud. But for my program I created my own word cloud analysis system.

To create a word cloud in R I need to use some libraries as require(tm), require(wordcloud). Those two libraries is needed for mining data and creating word cloud. I also need another library for visualization. Which is require (RColorBrewer). It will provide us the theme for word cloud visualization.

Then I cleared my existed data and removed unwanted data as punctuations and stop words. I converted words in lower case and converted them into plain text. Then a word will be considered as valid for word cloud. Higher scorer will be the largest word considering that all those words who have lower value will be smaller in size.

I wrote an R script to create word cloud for analyze the twitter data. The script is given below:

```
getwordCloud <- function (txtdataframe , inText , emotion ) {
  emos = levels(factor(txtdataframe$emotion))
  n_emos = length(emos)
  emo.docs = rep("", n_emos)
  txtTweets = removeCustomeWords(inText)

  for (i in 1:n_emos){
    emo.docs[i] = paste(txtTweets[emotion == emos[i]], collapse=" ")
  }
  corpus = Corpus(VectorSource(emo.docs))
  tdm = TermDocumentMatrix(corpus)
  tdm = as.matrix(tdm)
  colnames(tdm) = emos
  require(wordcloud)
  suppressWarnings(comparison.cloud(tdm, colors = brewer.pal(n_emos, "Dark2"),
}
```

Figure 4.5 Word cloud script

After running the following script I will get my expected output which is word cloud of my collected tweets. There is different themes for word cloud and they define what will be the color for a keyword based on its value. Here I used dark2 theme for my word cloud.

When I run this script I will get the word cloud. Word cloud size depend on the number of tweets and keywords. A demo output for word cloud is given below:

Jahangirnagar University
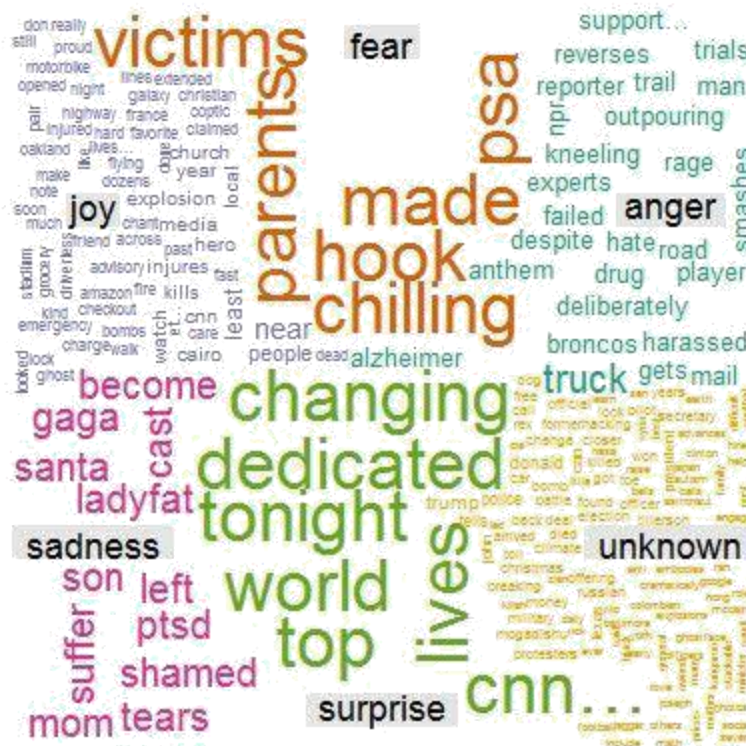
Figure 4.6 Word cloud for tweet data

From the following word cloud I can see that the output are divided in 6 different classes. They are joy, sadness, surprise, anger, fear and unknown. Each class represent its own type of emotions and frequent words for their tweets. Each class have its own color, so that user can understand which word belongs to which class.

4.1.4 SCORING

Calculating score came from the positive and negative scores of the tweets. Basically in this section I tried to show the number of tweets having same value for positive or negative sentiment. The higher the score, the stronger the sentiment. According to my consideration all the tweets cannot be counted as average positive or negative. Because

some of them are simple positive sentence and some of them are negative sentences. But there are also some sentences those can strongly express the sentiment as positive or negative.

So, this calculation was added to measure the strong sentiment state for the data I got from twitter. All those tweets who are having same sentiment value will be added to the same sentiment score bar.

For example, suppose I have 5 individual data or tweets:

1. I am having some trouble today.
2. That party way awesome, mind blowing, excellent. I just loved it.
3. Mango is a tasty Fruit.
4. Submit your CV in here http//…..
5. He is a good man.

Considering them from the 1st tweet I got the score -1, 2nd tweet I got the score 4, 3rd tweet I got the score 1, 4th tweet I got the score 0, 5th tweet I got the score 1.

So now here I got the score (-1, 4, 1, 0, 1), where 2 tweets are having score 1 and all others are having different scores. So I got the bar chart as,

Score	-1	0	1	2	3	4
Tweets	1	1	2	0	0	1

Figure 4.7 scoring system table

I did the same thing using R programming. I calculated the score and plotted them as bar chart. The R program for scoring output is given below:

Jahangirnagar University

```
65
66  hist.plot <- ggplot() + geom_histogram(data=sentiment.scores, aes(x=score))
67
68  hist.plot = hist.plot + ylab("Number of Tweets")
69  hist.plot = hist.plot + xlab("Tweets Having Sentiment score")
70  hist.plot = hist.plot + ggtitle("Sentiment Analysis of Tweet Data")
71
72  #hist.plot + theme_economist() + scale_colour_economist()
73  hist.plot + theme_solarized_2(light = FALSE) + scale_color_solarized("blue")
74
75  png(filename = "scores.png")
76  plot(hist.plot)
77  dev.off()
78
```

Figure 4.8 Scoring histogram plot / bar chart program

In the given program I also stored the generated histogram plot or bar chart in my working directory for further analysis. In the plot I set the row as "Number of tweets" and column as "Tweets having sentiment score". The score and the tweet numbers are calculated as I previously described. The output histogram plot or bar chart for scores are given below:
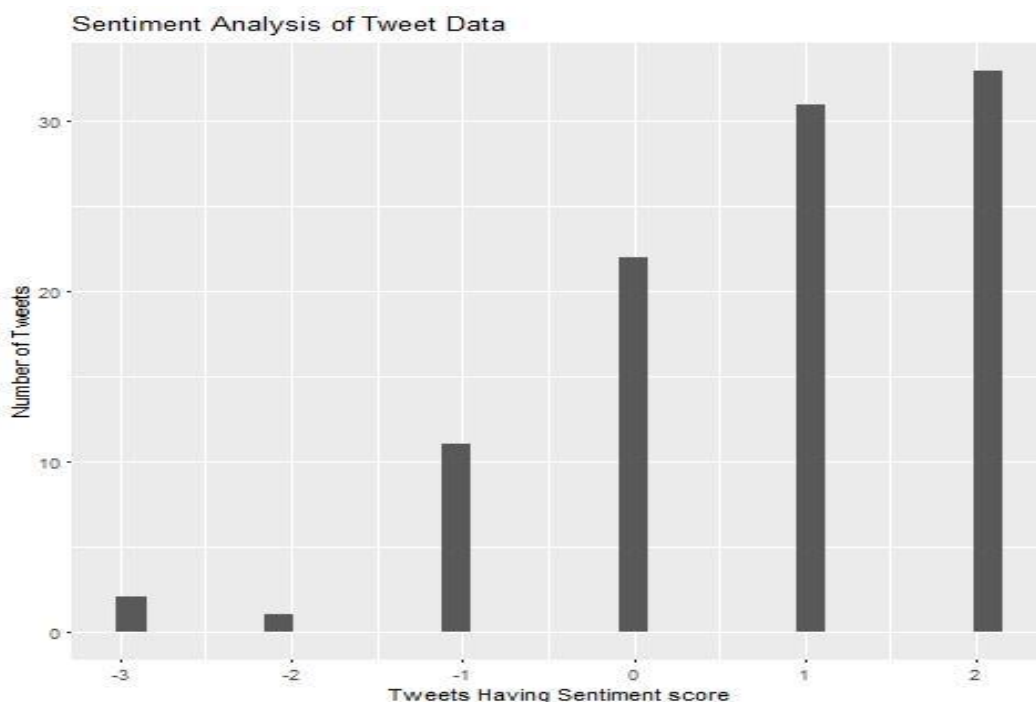

Figure 4.9 Score chart for tweets

Jahangirnagar University

Following figure is showing us the score dependent output based on my previously done analysis. Outputs will be different based on different time, search keyword, number of tweets I collected and tweet updates.

4.2 FINAL OUTCOME

After completing my analysis I wanted to create a user interface where general users will be able to analyze data. So I choose shiny as my tool to create the user interface. The main benefit of shiny is it can create web applications which can interact with R.

To build a shiny app two thing is required. One of them is ui, and another one is server. UI and server can be kept in same file like "app.r", or in separate files like "ui.r" and "server.r".

UI contain the outputs and the webpage. On the other hand, server contain the main program and inputs. So I kept my main program in the server file and created a webpage where I kept a form by which I can search anything I want to analyze. On another part of the page outputs will be shown.

To run the program I need to enter into the R software and go to the UI or server script. There will be a "Run App" button. If I press that the program will run into instructed window for example browser.
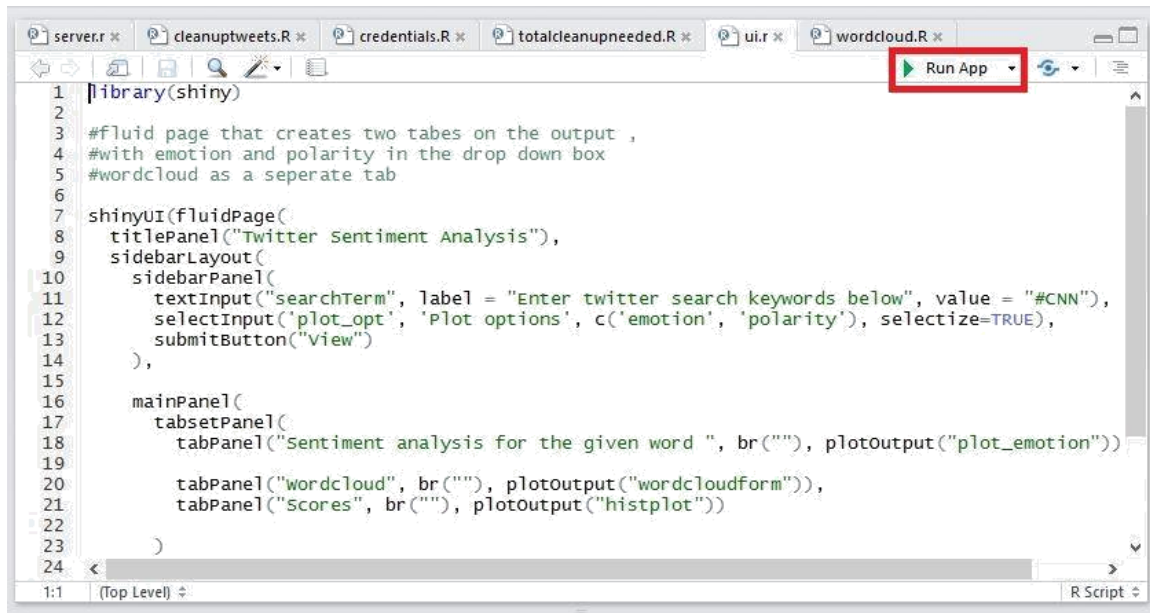
Figure 4.10 Running Application

When the webpage opens in the browser then I will be able to see a search box on the left. I need to insert whatever I want to search. Then I need to click on "VIEW" button to view the analyzed result.

Figure 4.11 Searching for tweets

Jahangirnagar University
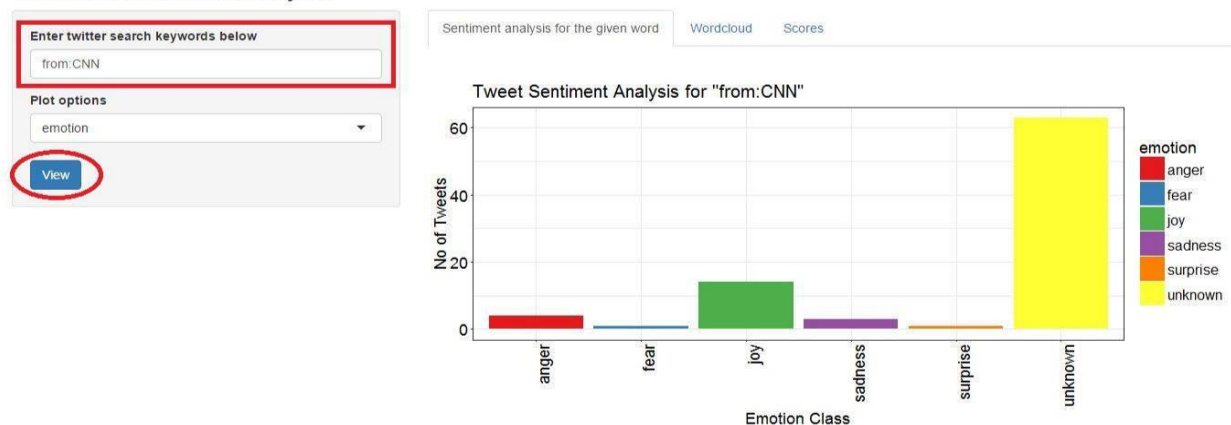
After that the whole R program will start on the background. It will collect tweets according to the search input. Then it will preprocess them. The preprocessed data will create an emotion class diagram and plot it into the "Sentiment analysis for given word" tab.

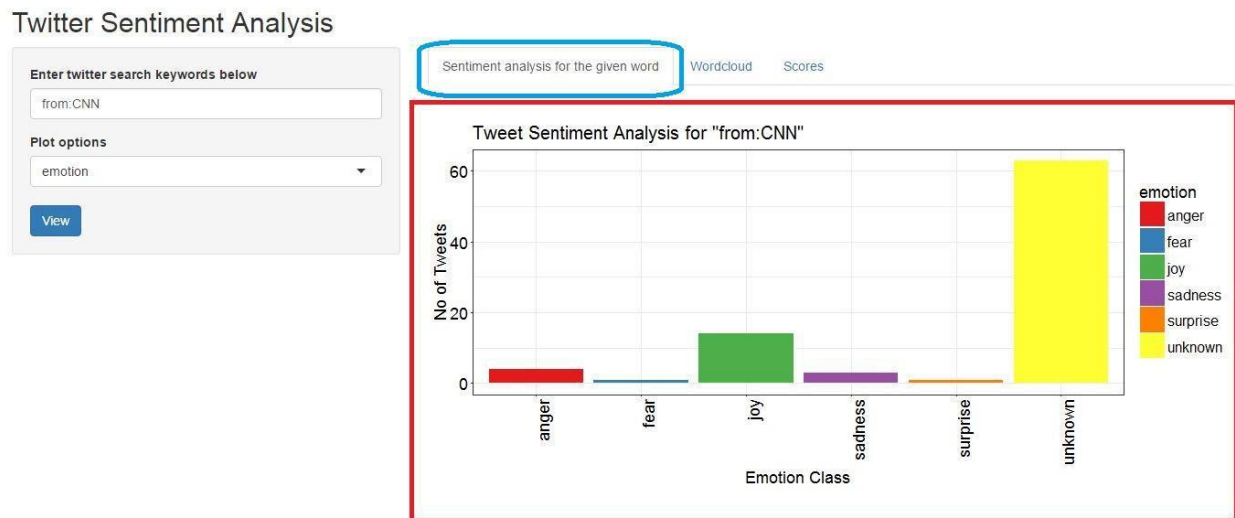

Figure 4.12 Emotion class output

If I want to view polarity class output I need to select "Polarity" from the form on the left side. Then the program will analyze the polarity and plot the diagram of polarity class instead of emotion class.
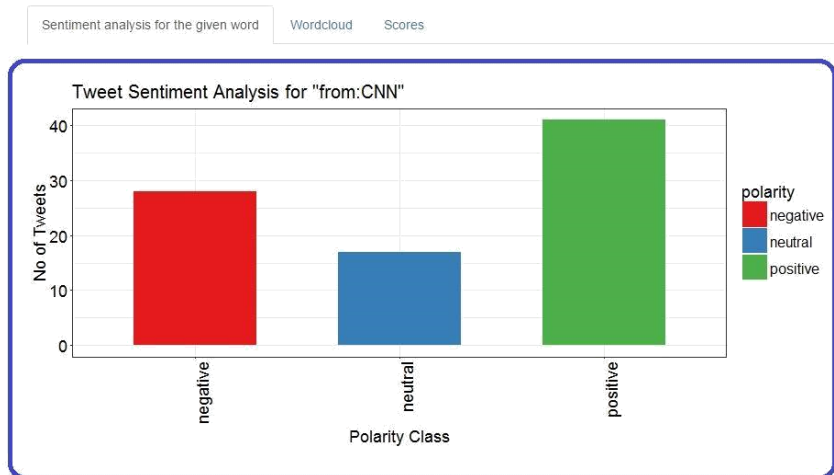
Figure 4.13 Polarity class output

To view the word cloud output I need to select on the "WORDCLOUD" tab and I will be able to see the word cloud in there. Depending on the processing speed this may take a while to view the output.

Figure 4.14 Word cloud output

My last part of output is score histogram plot. To view the scores histogram bar plot I need to click on the ''scores'' tab. Then the plot will appear in the web page.

Figure 4.15 Scores histogram output

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

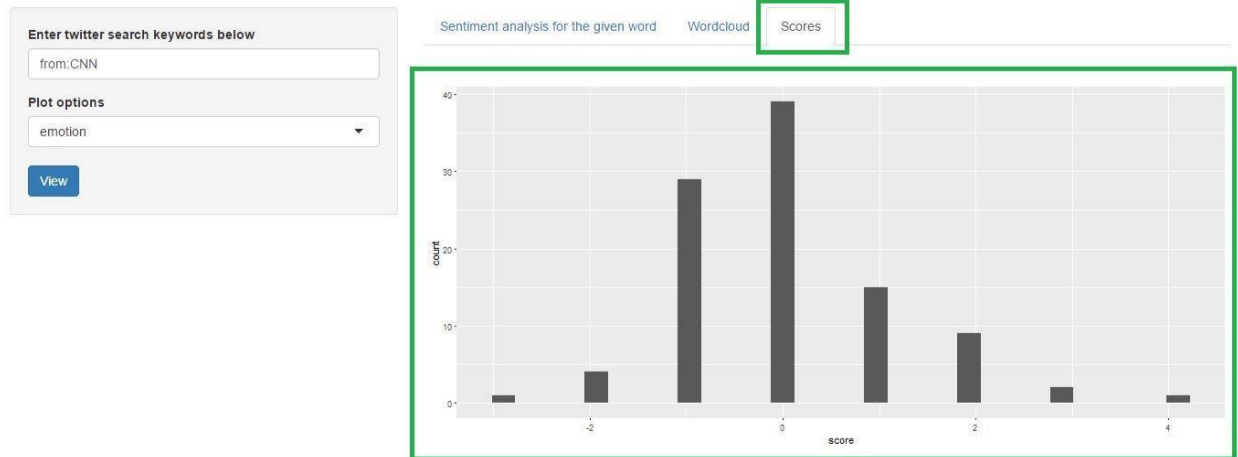As I can see that the output of this project is giving us a overall statistical view for selected or given search keyword. So that I can find out the emotional state for targeted keyword or data or person. Those outputs can be used for different sectors like education, research, business. I are still working on it to get better result and use it in specific work to help people. But learning and applying my methods for the future development is my main challenge.

5.1 CONCLUSION

In this project I tried to develop a complete project on sentiment analysis of tweet data. Which will be able to collect tweet data and analyze them and based on the analysis it will provide statistical sentimental structure of the search keyword. Even it will be easy for non-technical people to use it and examine the output. While I were working this project I learnt a lot of things, also faced a lot of challenges also. From this project I learnt data mining and analysis. Now I have real time data analysis experience. I enjoyed my project work.

On the other hand, the main challenge I faced is resource. There is a very few good websites for learning data mining. Even if I face any problem it is difficult to find solution for that. That's why I lost a lot of time in finding solution.

Though it took a lot of time to learn this step by step and all those steps were very small, but I learnt from those. For my analysis I learnt R platform and language. I also got idea about different packages inside it and how to work with them. I tried to not to make it complex and obtain a high efficiency result from this project.

5.2 FUTURE SCOPE

Sentiment analysis is already evolving from general (positive, negative and neutral) to much more complex or more granular and deep understanding. So, the demand of sentiment analysis is increasing in both side of research and business. Researchers are working on the accuracy of the algorithm and development of the lexicon based analysis. On the other hand, business companies are working on the market policy and customer satisfaction analysis to develop their business.

Out research work has potential of both to be used as commercial aspects or to do further research. For commercial aspect, business companies can find out their customer satisfaction based on tweets. So that they will be able to change their business policy to improve their benefits and attract new customers to their product. On the other hand, researchers can collect data for individuals to get the sentimental status of a person and use that data for further research. Even development of accuracy of algorithms are also can be done by my research work. In future I are planning to implement more algorithms in my research work to make it more accurate for sentiment analysis. I want to contribute more in this research field by keeping carry on study.

REFERENCES

1. SENTIMENT ANALYSIS; https://en.wikipedia.org/wiki/Sentiment_analysis; Last Accessed on 23th August 2016.

2. RELATED WORKS; https://www.quora.com/What-are-the-applications-of-sentiment-analysis-Why-is-it-in-so-much-discussion-and-demand; Last Accessed on 14th September 2016.

3. CHALLENGES; http://www.adweek.com/prnewser/5-key-challenges-in-sentiment-analysis/116604; Last Accessed on 20th September 2016.

4. CHALLENGES; http://www.sciencedirect.com/science/article/pii/S1018363916300071; Last Accessed on 20th September 2016;

5. Anais Collomb, Crina Costea, Damien Joyeux, Omar Hasan, Lionel Brunie, A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation, University of Lyon, INSA-Lyon, F-69621 Villeurbanne, France.

6. EMOTION CLASS; https://blog.hootsuite.com/beginners-guide-sentiment; Last Accessed on 11th October 2016.

7. EMOTIONCLASS;https://www.microsoft.com/developerblog/real-life-code/2015/11/30/ Emotion-Detection-and-Recognition-from-Text-using-Deep-Learning.html; Last Accessed On 11th October 2016.

8. EMOTION CLASS; http://betterevaluation.org/en/evaluation-options/wordcloud; Last Accessed on 11th October 2016.

9. WORDCLOUD; https://web.njit.edu/~da225/NetHelp/default.htm?turl=Documents%2 Fcurrentexamplesofsentimentanalysis.html; Last Accessed on 23rd October 2016.

10. RESEARCH METHODOLOGY; https://www.tracx.com/resources/blog/out-of-the-dark-ages-the-rise-of-social-media-sentiment-analysis-part-1-the-dark-ages; Last Accessed on 7th July 2016.

11. SENTIMENT ANALYSIS; http://searchbusinessanalytics.techtarget.com/definition/ opinion-mining-sentiment-mining; Last Accessed on 5th August 2016.

12. RESEARCH METHODOLOGY; http://www.slideshare.net/sumit786raj/sentiment-analysis-of-twitter-data; Last Accessed On 10th July 2016.

13. LEKHA R. NAIR, DR. SUJALA D. SHETTY, STREAMING TWITTER DATA ANALYSIS USING SPARK FOR EFFECTIVE JOB SEARCH, Journal of Theoretical and Applied Information Technology, 20th October 2015. Vol.80. No.2.

14. Vishal A. Kharde, S.S. Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications (0975 – 8887), Volume 139 – No.11, April 2016.

15. Penchalaiah.C, Murali.G, Suresh Babu.A, Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 8, October 2014.

16. Manjunath Mulimani, Nireeksha V Shetty, Nikshitha Shetty, Renita Maria Lolo, Rajashree L, Data Analysis on Current Affairs Using R, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, April 2015.

17. LEKHA R. NAIR, DR. SUJALA D. SHETTY, STREAMING TWITTER DATA ANALYSIS USING SPARK FOR EFFECTIVE JOB SEARCH, Journal of Theoretical and Applied Information Technology 20th October 2015. Vol.80. No.2.