SENTIMENT PREDICTION OF TWEET DATA

BY

Nusrat Orith Nishat
ID: 151-15-5015

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Nusrat Jahan
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Md Azizul Hakim Shuvo
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2018

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Nusrat Jahan, Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Nusrat Jahan
Designation
Department of CSE
Daffodil International University

Co- Supervised by:

_____

Md Azizul Hakim Shuvo
Designation
Department of CSE
Daffodil International University

Submitted by:

_____

Nusrat Orith Nishat
ID: -151-15-5015
Department of CSE
Daffodil International University

ACKNOWLEDGMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I really grateful and wish my profound my indebtedness to **Nusrat Jahan**, **Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Data mining*" to carry out this project. Her 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 the Almighty Allah and **Dr. Syed Akhter Hossain, Professor and Head,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

As descriptive as the title is, “Sentiment Prediction of Tweet Data” is a project that harnesses the power of modern information and technologies, various tools like R to analyzed people’s sentiment on a certain matter through their tweets. If developed properly in the future this project can help organizations as well as individuals to analyze and understand what people are thinking or feeling about an affair. In order to analyze data effectively one needs to have a solid understanding of data science and some other technical skills. That means someone who has never wrote computer programs or does not have a good mathematical knowledge will not be able to perform such analysis. Our goal was to make a software that makes this entire process of analyzing data and making a report more accessible to anyone. Although there are many tools and platforms available for data analysis, I went for R programming language. Because it is not only specialized for this task but also has a huge amount of resources online. So extending the software becomes easier in the future. I have finished working on my project. All it does is it loads up tweets from the internet base on a certain topic and analyses with Naive-Bayes algorithm. Later on displays reports based on several parameters in a human readable chart format.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Introduction

In the present age of information and technologies, information is power. The more information we have on a certain fact the more precise we can be with our decisions and deductions. We all know that information is nothing but processed data. Now data is not anything rare. Social networks like Facebook and Twitter where people are posting their thoughts, opinions on hundreds of matters can act as a gigantic pool of data. Now a days 500 million tweets are posted on the walls of twitter on a daily basis. So it has become very important to analyze them in order 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.

1.2 Motivation

Along with visual or audible data the importance of textual data has grown drastically in recent years. In fact it seems like that textual data has become a lot more crucial than other forms. People share their opinions and sentiments on the internet every day. I most cases these opinions are very real and not made up. Opinions are a usually subjective expression that describes person's sentiment, feelings towards the object or service.

One of the most famous social networks of present time is Twitter. Twitter lets it's user to express their thoughts in the form of micro blogs. These micro blogs are nothing but text data consisting of 280 Unicode characters. By analyzing these data i can decide whatever the user's mental situation is going on. I can even find out what is the most talked about topic in a continent within a certain period of time. I realized that it can help people in various ways. Maybe organizations can necessary steps to deal with people's sentiment on a matter. This idea of predicting the near future through people's thinking worked as the motivation behind my project.

1.3 Rational Of the Study

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 Analytic.
- 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.

1.4 Research Questions

In this research some questions may raise that-

- Is it really possible to predict sentiment from text?
- Can those reports be used for other purpose?
- What are the main applications of this research?

For all those questions it can be said that almost every kind of post uploaded in twitter have some sentiment related words. Using them it is possible to predict the sentiment. On the other hand this research can be used in market analysis, decision making, user review analysis and so on.

1.5 Expected Output

I decided to work on sentiment analysis 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. Building an app which will be able to collect twitter data for example tweets and analyze them was my target. After the analysis the application will provide a statistical overview of sentimental situation of targeted people.

Though there are a plethora of languages and tools in the market I decided to make my application using R programming language. Where I will be able to search random data by providing a limit number.

1.5 Report Layout

There are five chapters in this research paper. They are: Introduction, Background, Research Methodology, Results and Discussion, Conclusion and Future.

Chapter one: Introduction, Motivation, Rationale Of the study, Research Question,

Expected Outcome, Report layout.

Chapter two: Related works, Research Summary, Scope of the problem, challenges.

Chapter three: Research Subject and Instrumentation, Data Collection Procedure, Statistical Analysis, Experimental layout.

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

Chapter five: Summary; Conclusion, Future Scope.

CHAPTER 2

Background

2.1 Introduction

Sentiment prediction is a computational process to classify human emotion in different classes. Sentiment prediction is possible for text, audio, video or image. But each of them will require different way, approach and algorithm. As I worked on text analysis, so this paper is containing data about only text mining. But if I want to start sentiment prediction, at first i need to know about the sentiment prediction and its approaches. Sentiment can be considered as feeling, attitude, emotion or opinion. Computationally identifying and processing some text, I can do Sentiment prediction to determine whether the writer's emotion on a particular topic or a subject or a product or a person is positive, negative or neutral. So using natural language processing categorizing, classifying or statistically viewing some analytical report from a bunch of text is generally considered as sentiment analysis.

I can define sentiment analysis in main three levels [5]. They are:

1. Document level sentiment analysis
2. Sentence level sentiment analysis
3. Phrase level sentiment analysis

Here I am going describe those levels of sentiment prediction:

Document level sentiment prediction: In the document level sentiment prediction I will consider a single document containing a single topic. So this level of sentiment prediction is not applicable for analyzing different forum or blog sites. The main challenge for document level analysis is there can be some texts that are not relevant to the topic. So whenever I apply this, we have to remove irrelevant sentences. For this level prediction I can use both supervised and unsupervised learning.

Any supervised learning algorithm like naïve bays [6] or vector machine and unsupervised learning algorithm like clustering or k-means algorithm can be used for this classification.

Sentence level sentiment analysis: In the sentence level sentiment analysis we will consider every sentence in the document. So this level of sentiment analysis is applicable for every kind of document as general document or forum or blog sites. Here by determining the positive and negative words we will be able to determine if the sentence is a positive or a negative sentence. Sentence level sentiment prediction is not applicable for the complex sentences. Like document level sentiment prediction, we can apply both supervised and unsupervised learning in sentence level sentiment prediction.

Phrase level sentiment prediction: In the phrase level sentiment analysis i will consider opinion or sentiment related words. This level of predict is a pinpointed approach for the sentiment prediction. In some cases exact opinion can be extracted on a specific topic from the text. But this level analysis faces the problem of negation and long range dependencies. Those words who appear near to each other are considered as a phrase in here.

Between all of them sentence level analysis is applied for this research. As sentence level analysis can be used for both supervised and unsupervised learning, so use of both supervised and unsupervised learning gives better result of the analysis.

Sentiment prediction is mainly important for business companies, political groups and different social organizations. By collecting customer review and satisfaction business companies can determine their next business strategy and success of their business. Political groups can manipulate people using current statistics and most discussed topics, so they will be able to know what they want from them. On the other hand, different organizations will be able to determine the support for them for a particular task or their work. Those are some examples for the importance of sentiment analysis. So we can say that by analyzing sentiment we can be beneficial in our practical life.

Modern era brought a very large field for computer scientists to work with data for determining sentiment of people. All around the world a lot of people are working on sentiment prediction right now. Continuous research brought us new features and more accurate result. Research done by Pang and Lee [7] brought so many different approaches like detection of human expression; classifying sentences as positive,

Negative or neutral; detection of subjective and objective sentences; classifying human emotion in different classes like anger, happy, sad etc.; application of sentiment prediction in different sectors [8].

Hatzivassiloglou and McKeown [9] and Esuli and Sebastiani [10] worked with the polarity detection from the phrases. Yu and Hatzivassiloglou [11] and Kim and Hovy [12] worked on the character limitation and determined that the twitter message analysis is much similar to the sentence level sentiment analysis [13]. But at recent time so many people are working on different topics of sentiment prediction like twitter, facebook, newspaper, blogs, novels etc. For example, Safa Ben Hamouda and Jalel Akaichi [14] worked on sentiment classification from Facebook for Arabic spring era. Zhaoxia WANG, Victor Joo Chuan TONG and David CHAN [15] was working on issues of data analysis of social media with new algorithms; Jonathan Bright, Helen Margetts, Scott Hale and Taha Yasseri [16] worked on the use of social media for research. Mika Viking Mäntylä, Daniel Graziotin and Miikka Kuutila [17] worked on evolution of sentiment analysis. S Padmaja, Prof. S Sameen Fatima and Sasidhar Bandu [18] worked on evolution of sentiment analysis and negation in newspaper.

All those researchers worked on those topics based on supervised and unsupervised learning. Even some cases hybrid methods are applicable for sentiment analysis. Application of different algorithms makes sentiment analysis process easier.
Sentiment analysis based on text faces different kind of difficulties. According to Professor Bing Liu from University of Chicago department of computer science, being sure about the accuracy for the analysis is difficult and it depends on the level of analysis, number of data sets, measurement and so on [19].

One of the most common problem is different meaning for the same text. There are some other problems as well like different languages, shortcut words written by the writer, typing mistakes etc. So while working with those type of problems it is really difficult to decide how to solve them [20].

On the other hand, negation words or sentences like ‘it is not good’ can create problems for algorithms to find out accurate result to determine whether it is a positive or a negative word by showing a positive word as negative and a negative word as positive.

2.2 Related Works

Sentiment prediction 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 – Semantic 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 now a days 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 prediction. Using sentiment analysis they determine marketing strategy, improve campaign success, improve product messaging and improve customer service.

Some sentiment prediction related works are:

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

5. Sentiment Prediction 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 prediction, Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao.
7. Decision Making Using Sentiment prediction from Twitter, M.Vasuki, J.Arthi, K.Kayalvizh.
8. All those are the recent research works on sentiment prediction. They are focused on different topics like customer review, decision making, NLP, twitter data, opinion mining etc.

2.3 Research Summery

Sentiment Prediction also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression

1. Polarity: if the speaker express a positive or negative opinion,
2. Subject: the thing that is being talked about,
3. Opinion holder: the person, or entity that expresses the opinion.

Currently, sentiment prediction is a topic of great interest and development since it has many practical applications. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media. With the help of sentiment prediction systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service.

Sentiment prediction 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 prediction that can become a valuable skill.

When I perform sentiment prediction on some content, the main searching point is the opinions in content and we 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 prediction 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.

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 prediction 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 prediction 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 prediction are endless. More and more we're 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.

2.4 Scope of the Problem

Sentiment analysis is one of the most popular topic of research all over the world, because till now though a lot of work has been done but still there is a lack of accuracy and understanding for the machine. Therefore people are working hard on the algorithm, sentiment word library and different techniques. In this research we are using an algorithm based technique known as Naïve Bayes algorithm. Our focus will be the improvement of the accuracy.

2.5 Challenges

One of the prime issue of sentiment prediction now is the accuracy. Several technical or conceptual 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 prediction 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 prediction try to detect the mental situation of a person. But sometimes it become tough to tell what the person meant. If we 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 now a day.

CHAPTER 3

Research Methodology

3.1 Introduction

In this chapter i will do the majority of my practical work. In order to start my work i need some data to work on. As i have started already i will use twitter data as our data as our data set. For collecting data i will use custom R script. This script will make contact with my application on twitter server and get tweet data by utilizing twitter's stats less application programming interface. Once the data has been collected it will preserved in unprocessed raw format. In order to get somewhat accurate result from my program i need to pre-process the data and make it usable for our purpose. Once pre-processed i can then run my program on the data set to determine results based on several parameters.. Last part of this step is concerned about the experiment layout. This experiment layout. This experiment layout will later serve as the model for our final product.

3.2 Research Subject and Instruments

As i selected sentiment prediction of tweet data as my project. So i needed to collect data from twitter. Everyday twitter generate 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 needed to consider this as my challenge for sentiment prediction.

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

I selected R as our 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 I started 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 DATACAMP website It is very helpful website because its shows how to learn R in step by step process.

After that i created a twitter app for data collection and using R language we created an application. This application can connect to twitter using internet and collect data and analyze them. Collected data are stored in a CSV file.

3.3 Data Collection Procedure

To collect data i needed a twitter application, which can give us 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 created an application named Sentiment Analysis BDA.

- **Creating Twitter App**

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

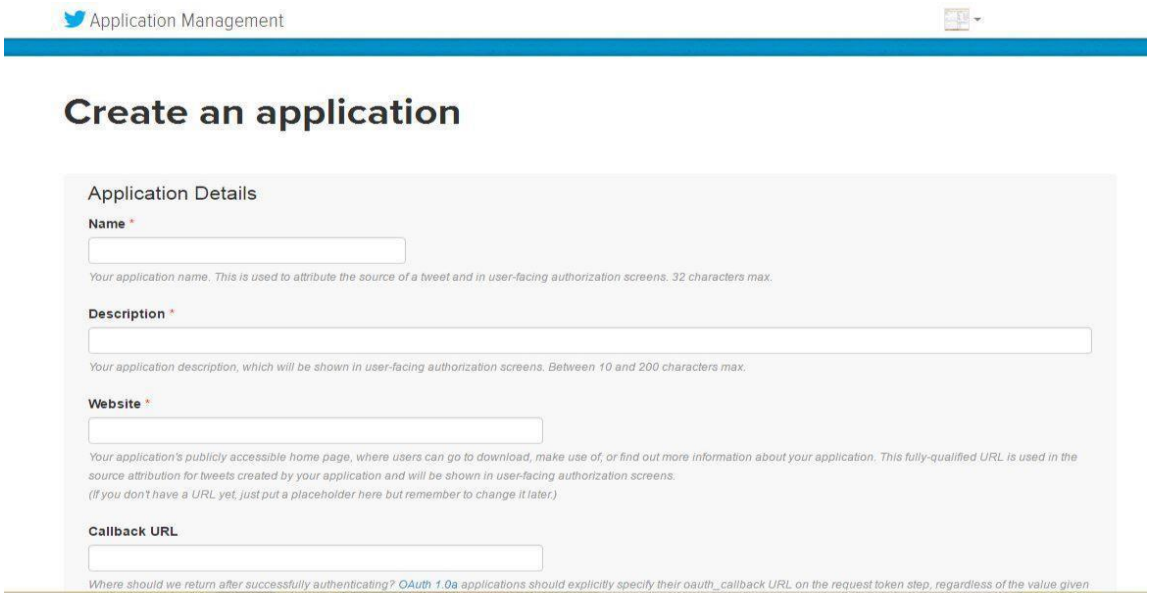

Figure 3.3.1: Creating a twitter application

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

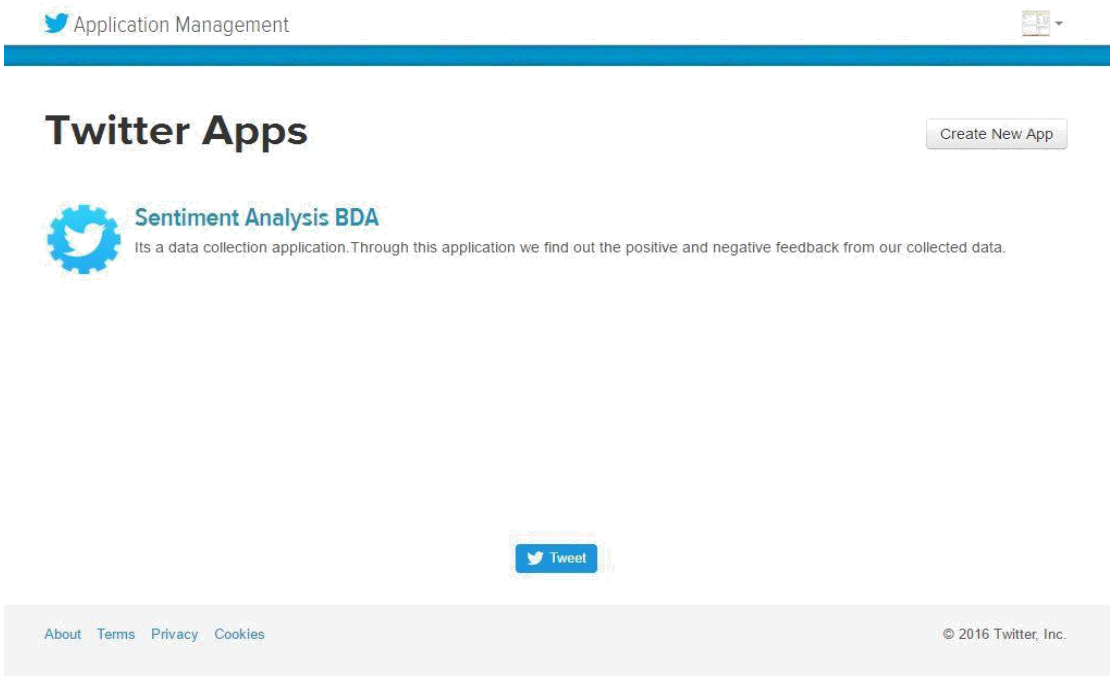

Figure 3.3.2: Application created in twitter

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

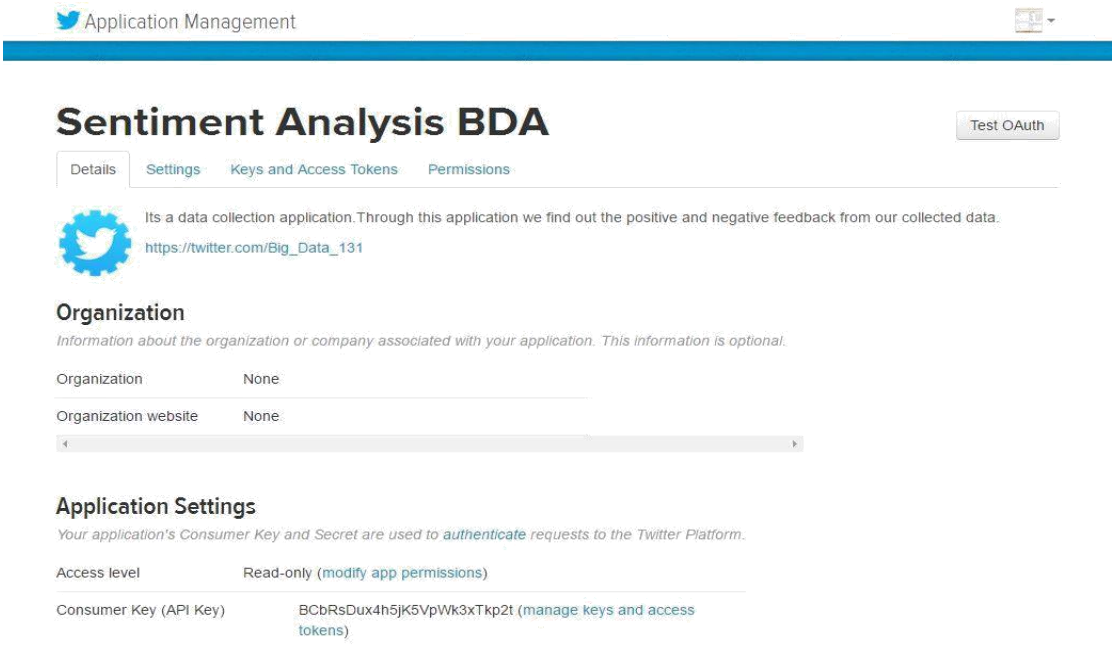

Figure 3.3.3: Application home page

- **R Setup**

When creating twitter application is done now i am ready to build our R application to collect data and store them. To create R application first i downloaded and installed the R application to our computer.

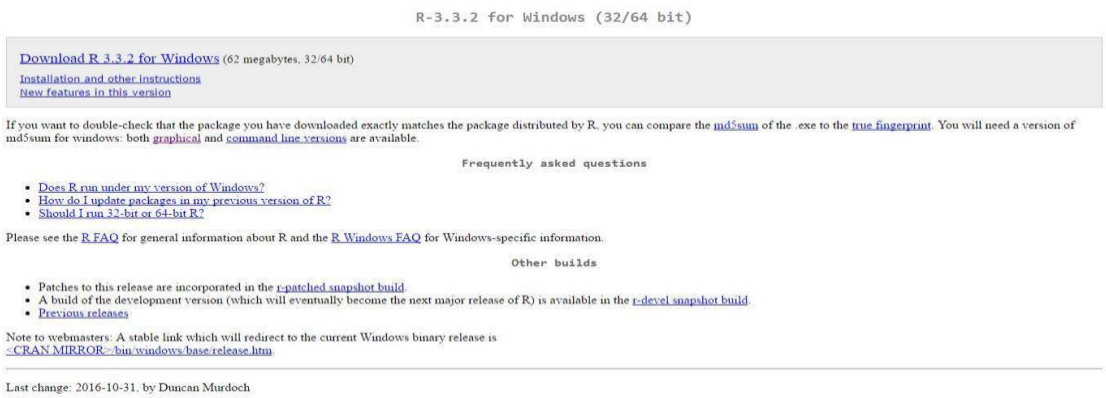

Figure 3.3.4: R download

After downloading and installing the R app i will get a user interface shown below:

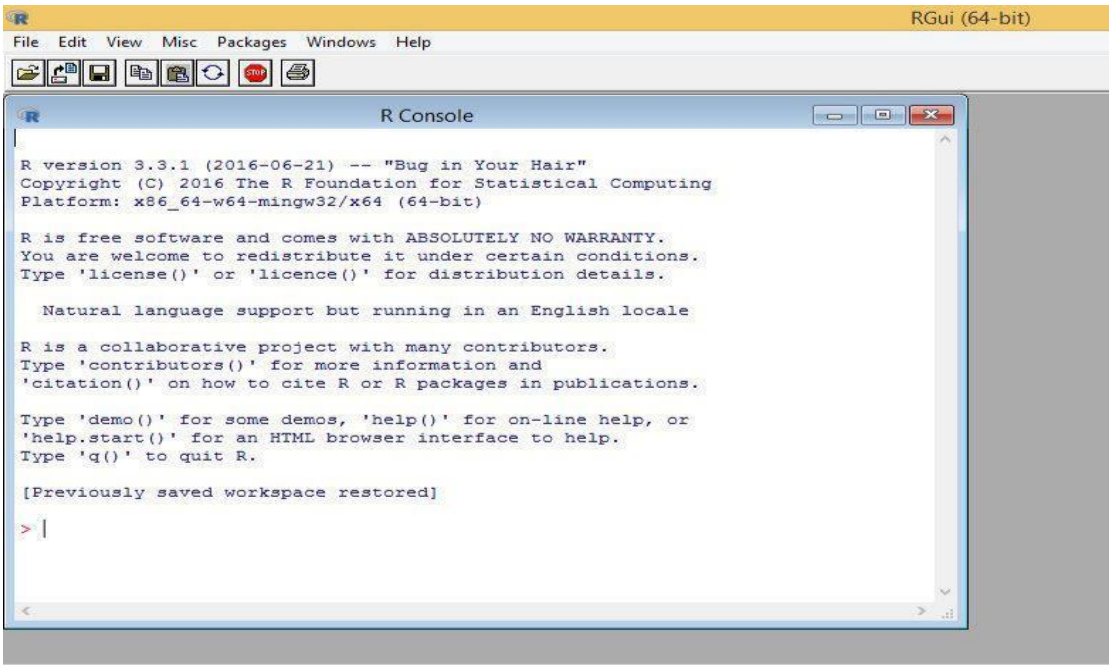

Figure 3.3.5: R user interface

But the user interface don't look that much comfortable to use and there is no visualization option. So now i have to install another application named R Studio.

R Studio 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.

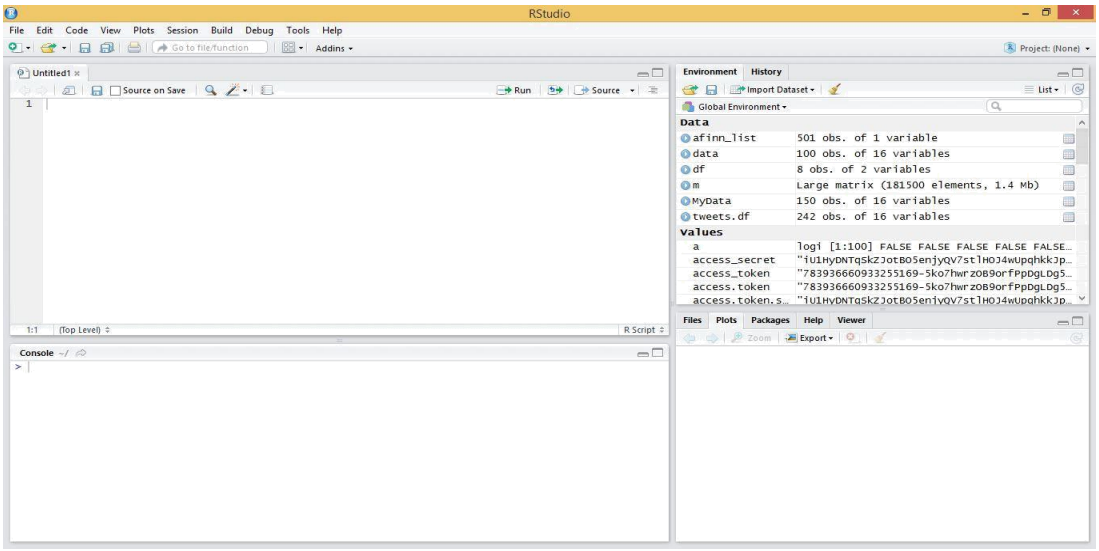

Figure 3.3.6: R Studio user interface

After installing necessary components now I am ready to write the program to collect tweets and store them.

- **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:

Table 3.3.1: Demo data Collection

text	favourite	favoriteCoureolyToSN	created	turncated	replyToSID	ID	reply to SID
She learned	FALSE	343 NA	9/30/18	FALSE	Na	1.05E+18	NA
An estimate	FALSE	158 NA	9/30/18	TRUE	NA	1.05E+18	NA
A Newly dis	FALSE	474 NA	9/30/18	TRUE	NA	1.05E+18	NA
Obamacare	FALSE	318 NA	9/30/18	FALSE	NA	1.05E+18	NA
A long- lost	FALSE	177 NA	9/30/18	TRUE	NA	1.05E+18	NA
Shockwave	FALSE	207 NA	9/30/18	TRUE	NA	1.05E+18	NA
Watch a rar	FALSE	466 NA	9/30/18	TRUE	NA	1.05E+18	NA

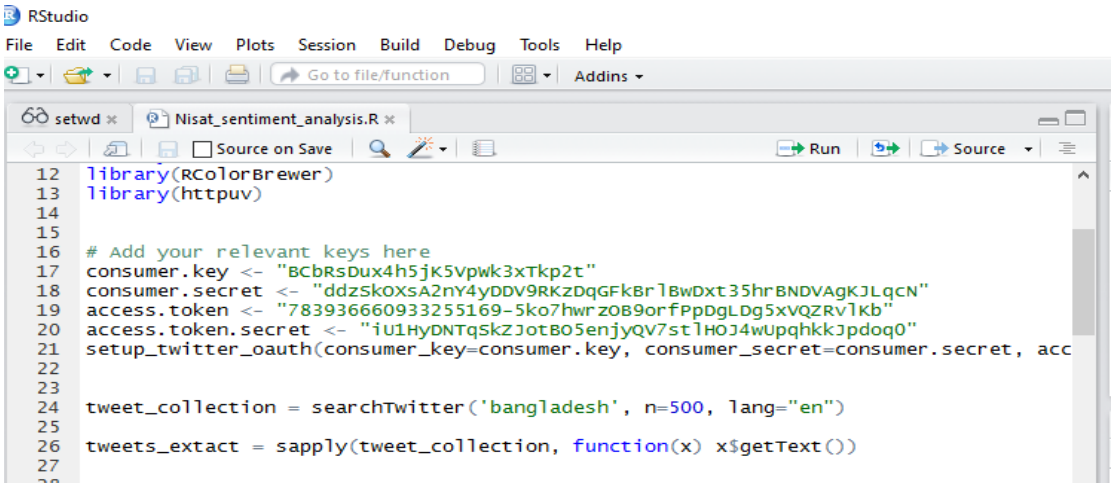

Figure 3.3.7: Data collection and store

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.

3.4 Statistical Analysis

When all the data are collected, our main task then starts. Which is “Sentiment prediction 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 our expected output of our project

Table 3.3.2: Proceeds Data

1	score	text
2	-1	She learned to swim at age 50 and now has 271 medals. Oh, and she's b
3	-1	An estimated 80,000 Americans died of flu and its complications last wir
4	0	A newly discovered dinosaur species that's believed to have weighed 26,
5	1	Obamacare is likely about to get a little cheaper for lots of people https:/
6	0	A long-lost warship has been discovered off the coast of China after an e
7	0	Shockwaves from the massive bombing raids during World War II reache

- **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:

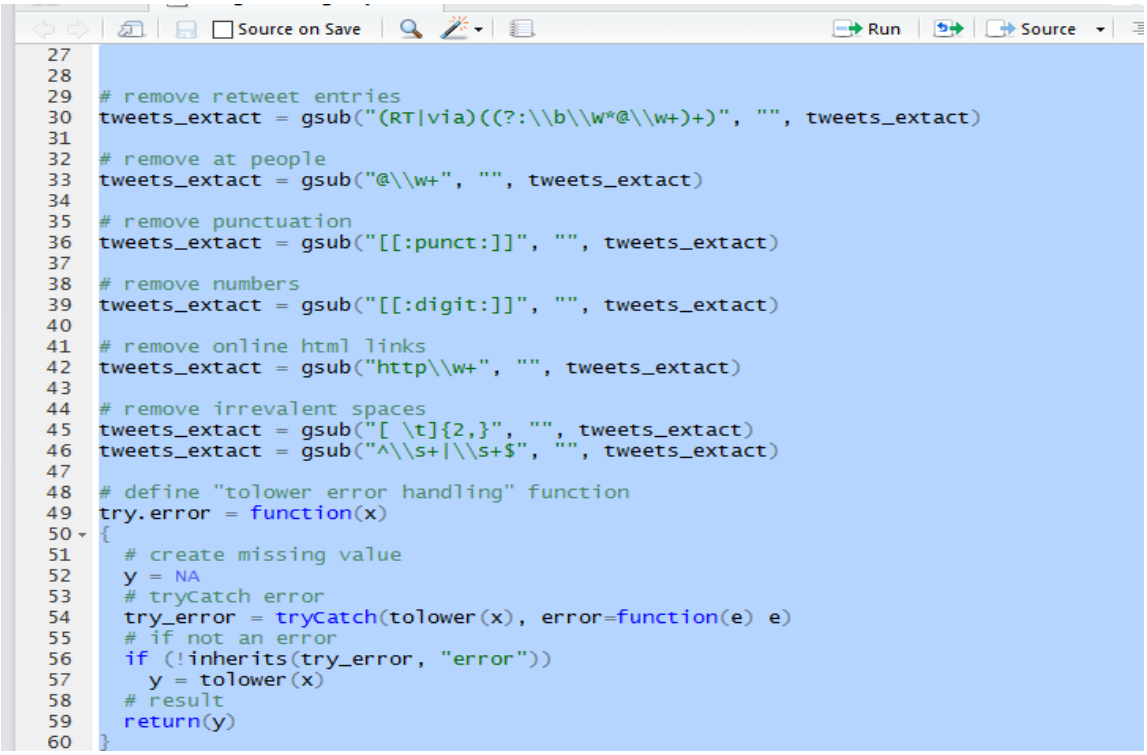

Figure 3.3.8: sorting data

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.

- **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 Naïve Bayes and Learning based approaches for concept-level sentiment analysis. Here I first divided my tweet sentences into words. So that i able to compare with our 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:

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;
; Opinion Lexicon: Negative

;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

2-faced
2-faces
abnormal
abolish
abominable
abominably
abominate
abomination
abort
aborted
aborts
abrade
abrasive
abrupt
abruptly
```

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;
; Opinion Lexicon: Positive
;

;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

a+
abound
abounds
abundance
abundant
accessable
accessible
acclaim
acclaimed
acclamation
accolade
accolades
accommodative
accomodative
accomplish
accomplished
```

Figure 3.3.9: 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 our previously stored positive and negative words which I am considering as my library for our 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.

- **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.

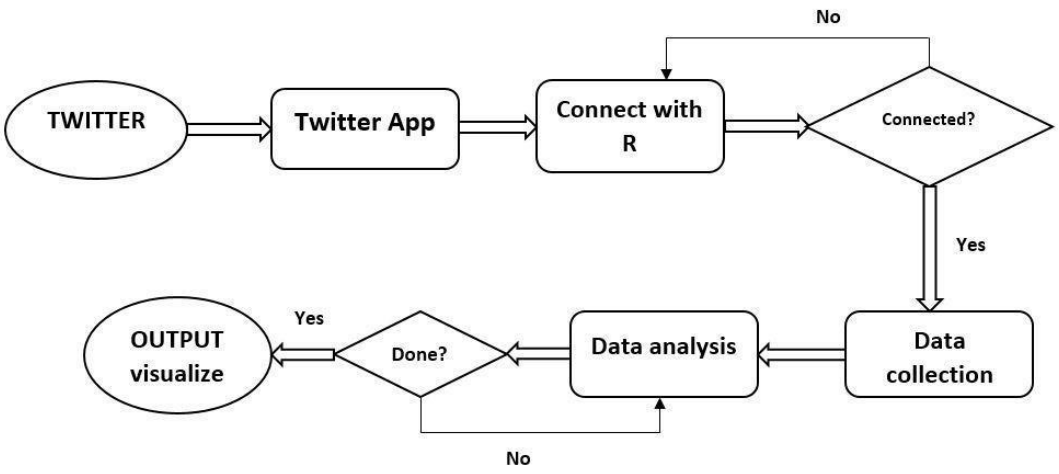

Figure 3.3.10: Data flow

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.

- **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 me 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 am 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 we designed a layout for our project. The layout is given below:

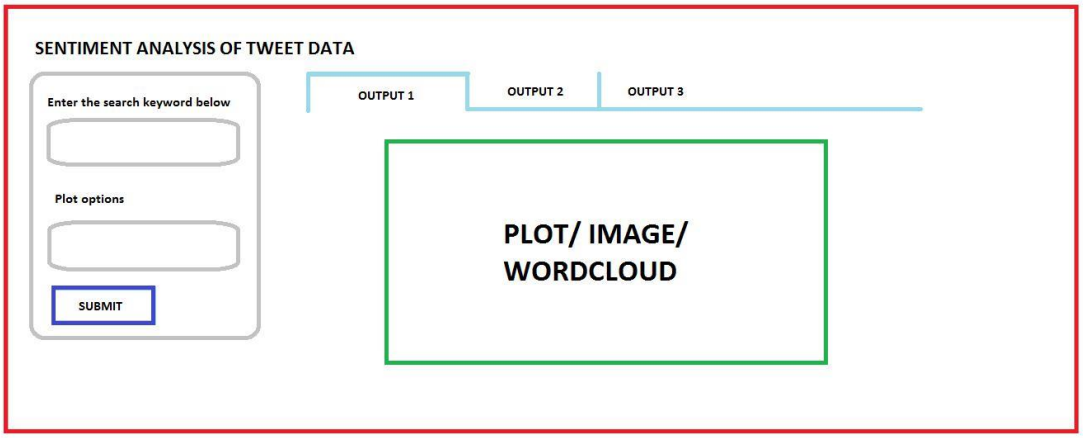

Figure 3.3.11: Experimental Layout

3.5 Implementation Requirements

To implement this research few things must be done first. Before starting the programming part a twitter account needed to be created. Later from that account an API account can be created. There will be an option to create a twitter application. If i just follow the instruction as i discussed before about creating application in twitter i will get some access token keys. Now my twitter part is ready. After that R software need to be installed and based on what I am going to do in the project required libraries needed to be installed in R platform. Than the system is ready to implement the project.

CHAPTER 4

Experiment Results and Discussion

4.1 Introduction

To use an application creating a user interface is the best option. So i created a user interface for my application so that anyone can use them easily. I used algorithms as Naive Bayes on the other hand i got four kind of data as output. All the outputs are visualized in our application via web browser.

4.2 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 3 different steps.

1. Emotion
2. Polarity
3. Word cloud

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 bar charts.

4.3 Descriptive Analysis

In this section descriptions about all those three types of analysis is discussed briefly. For reminder all of those are analyzed using the naïve bays algorithm in the R platform and R was used because this is one of the best tool for statistical analysis.

- **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 prediction 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 the emotion classes. Between all of them unknown class was calculated by the unknown tweets number.

To get the emotion class we need two libraries. They are require (play) and require (stringer). Where plyr is a tool for Tools for splitting, applying and combining Data and stringer 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 4.4.1: Emotion class pseudo 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.

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:

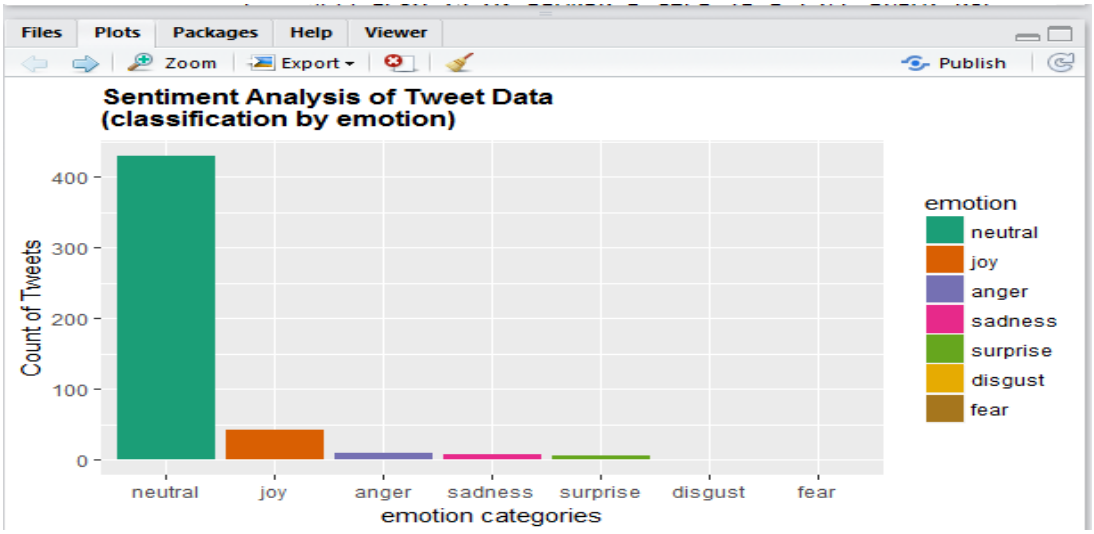

Figure 4.4.2: Emotion class bar diagram

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.

- **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.

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 are given below:

Table 4.4.1: Positive And Negative word

Positive words	Abound, abounds, acclamation, accolade, accolades, 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 bag of words are used to find out if the tweet is positive or negative. For the each positive or negative word every tweets will get 1 score.

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 contain tweet value 0, it will be under neutral tweet. Counting all of them we will get a proper chart of positive, negative and neutral tweets.

To calculate the polarity of the text i wrote 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 4.4.3: Polarity class pseudo 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:

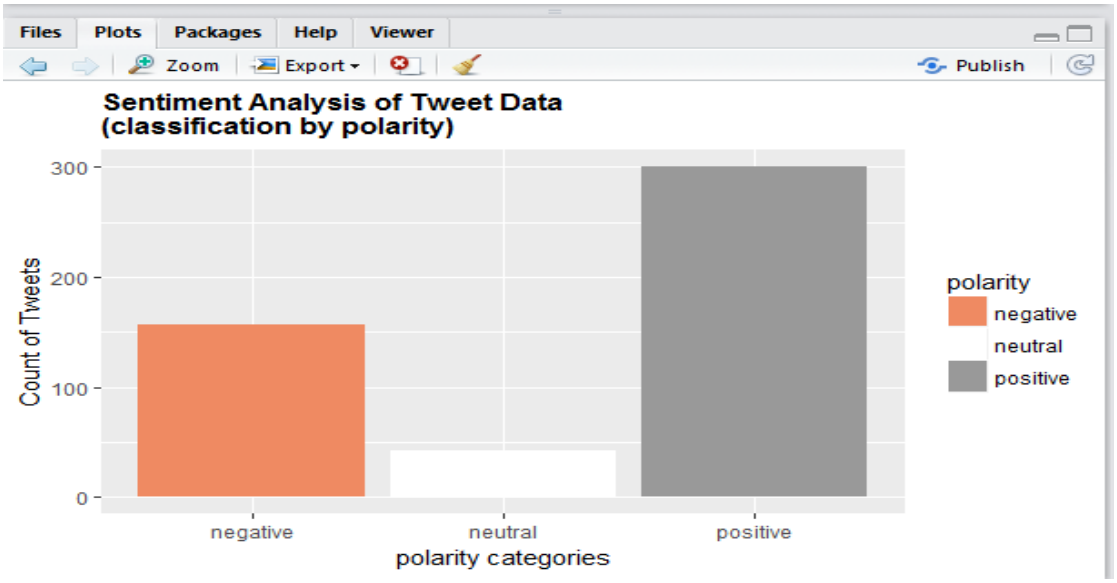

Figure 4.4.4: Polarity class bar diagram

- **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 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 our document file or the website link and they provide us the word cloud. But for my program I created our own word cloud analysis system.

To create a word cloud in R i need to use some libraries as require(tm), require (word cloud). Those two libraries is needed for mining data and creating word cloud. We 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 4.4.5: Word cloud script

After running the following script i will get our 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:

Figure 4.4.6: 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.

4.4 Summery

Though this may see complex to review a keyword using only one chart but all of them together can create clear picture on a topic. On the other hand as all those charts are dependent on each other, so it is also a plus point for the user of this application. As if i consider an emotion for example anger and see the corresponding word cloud keywords it will be easy to say about what users are feeling angry. If i consider another part, i will notice that using the polarity chart it is possible to verify if the emotion chart values are correct and giving us accurate answer.

CHAPTER 5

Conclusion and Future Scope

5.1 Summary of the Study

As I can see that the output of this project is giving me 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 still working on it to get better result and use it in specific work to help people. But learning and applying our methods for the future development is my main challenge.

5.2 Conclusions

In this project I tried to develop a complete project on sentiment prediction 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 was 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.

5.3 Implication for Further Study

Sentiment prediction is already evolving from general (positive, negative and neutral) to much more complex or more granular and deep understanding. So the demand of sentiment prediction 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 our research work. In future I am planning to implement more algorithms in our research work to make it more accurate for sentiment prediction. I want to contribute more in this research field by keeping carry on study.

REFERENCES

[1] Krishna, D. S. Kulkarni, G A Mohan A.: Sentiment Analysis-Time Variant Analytics. In international Journal of Advanced Research in Computer Science and Software Engineering, ISSN. 2277 128X, vol. 5, Issue. 3 (2015).

[2] Celikyilmaz , A., Hakkani-Tur, D. Feng J.: Probabilistic model based sentiment analysis of twitter messages. In: Spoken Language Technology Workshop (SLT), pp. 79-84, 2010 IEEE (2010).

[3] Muhammad, I., Yan, Z.: SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. In: International Journal of Digital Curation, pp. 133, DOI: 10.21917 (2015).

[4] Sashays R., Abraham, A.: Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. In: International Journal of Advanced Research in Artificial Intelligence, vol. 2, no.2 (2013).

[5] Varghese, R., Jayasree, M.: A SURVEY ON SENTIMENT ANALYSIS AND OPINION MINING. In: International Journal of Research in Engineering and Technology, EISSN. 23191163, PISSN. 2321-7308, vol. 2 (2013).

[6] Naïve Bayes for Machine Learning, <<http://machinelearningmastery.com/naive-bayes- for-machine-learning.>>

[7] Pang B., Lee, L: Opinion mining and sentiment analysis. In: Foundation and Trends in Information Retrieval, vol. 2, Nos. 1–2, DOI: 10.1561/1500000001, pp. 1-135 (2008).

[8] Svetlana, K. Zhu, X. Mohammad, S.M: Sentiment Analysis of Short Informal Texts Svetlana. In: Journal of Artificial Intelligence Research, vol. 50, pp. 723–762 (2014).

[9] Hatzivassiloglou, V., McKeown, K.R.: Predicting the Semantic Orientation of Adjectives. In: Eighth conference on European chapter of the Association for Computational Linguistics archive, pp. 174-181, Madrid, Spain (1997).

[10] Esuli, A., Sebastian, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceeding of the 5th Conference on Language Resources and Evaluation, Genoa, Italy (2006).

[11] Yu , Hatzivassiloglou, Y.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan (2003) .

[12] Kim, S.M., Hovey, E.: Determining the sentiment of opinions. In: 20th international conference on Computational Linguistics, no. 1367, Geneva, Switzerland (2004) .

[13] Kouloumpis, E., Wilson, T., Moore, J.: Twitter Sentiment Analysis: The Good the Bad and the OMG! In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 538-541, Barcelona, Spain (2011) .

[14] Hamouda, S.B., Akaichi, J.: Social Networks' Text Mining for Sentiment Classification: The case of Facebook' statuses updates in the "Arabic Spring" Era. In: International Journal of Application or Innovation in Engineering & Management, vol. 2, Issue. 5, ISSN 2319 – 4847 (2013) .

[15] Wang, Z., Tong, V.J.C., Chan, D.: Issues of social data analytics with a new method for sentiment analysis of social media data. In: IEEE 6th International Conference on Cloud Computing Technology and Science, EISBN. 978-1-4799-4093-6, PISBN: 978-1-4799-4092-9 Singapore, Singapore (2014) .

[16] Bright, J., Margetts, H., Hale, S., Yasseri, T.: The use of social media for research and analysis: a feasibility study. In: Department for Work and Pensions, ISBN 978-1-78425-407-0, London, England (2014).

.

[17] Mäntylä, M.V., Graziotin, D., Kuutila, M.: The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers (2016).

[18] Padmaja, S., Fatima, S.S., Bandu, S.: Evaluating Sentiment Analysis Methods and Identifying Scope of Negation in Newspaper Articles. In: International Journal of Advanced Research in Artificial Intelligence, vol. 3, no.11 (2014).

[19] What are the applications of sentiment prediction? <<https://www.quora.com/What-are-the applications-of-sentiment-prediction.>>.

[20] Vohra S. Teraiya, J.: Applications and Challenges for Sentiment Analysis: A Survey. In: International Journal of Engineering Research & Technology, e-ISSN: 2278-0181, Vol.2, Issue. 2 (2013).

[21] Why use the R Language? <<http://www.burns-stat.com/documents/tutorials/why-use-the-r language. >>

[22] What Are Emotion and the Types of Emotional Responses? <<https://www.verywell.com/whatare-emotions-2795178. >>

[23] Kharde V., A.: Sentiment Analysis of Twitter Data: A Survey of Techniques. In: International Journal of Computer Applications, Volume 139, No.11 (2016).

[24] Karanasou M., Ampla A. Doulkeridis C. Halkidi M.: Scalable and Real-time Sentiment Analysis of Twitter Data. In: IEEE 16th International Conference on Data Mining Workshops, ISSN: 2375-9259, Barcelona, Spain (2017).