

# **MARKET EVALUATION USING LEXICON BASED SENTIMENT ANALYSIS**

**BY**

**Shifun Neher Mofida ID: 151-15-4984**

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

Supervised By

**Dr. Md. Ismail Jabiullah**

Professor

Department of Computer Science and Engineering  
Daffodil International University



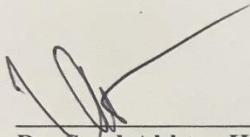
**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH DECEMBER 2019**

## APPROVAL

This Project titled “Market Evaluation Using Lexicon Based Sentiment Analysis”. submitted by Shifun Neher Mofida, ID No: 151-15-4984 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on December 5, 2019.

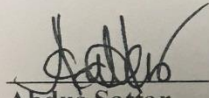
## BOARD OF EXAMINERS



**Dr. Syed Akhter Hossain**  
**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

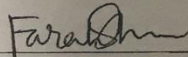
**Chairman**



**Abdus Sattar**  
**Assistant Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

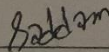
**Internal Examiner**



**Farah Sharmin**  
**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



**Dr. Md. Saddam Hossain**  
**Assistant Professor**

Department of Computer Science and Engineering  
United International University

**External Examiner**

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Md. Ismail Jabiullah, Professor, Department of CSE Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

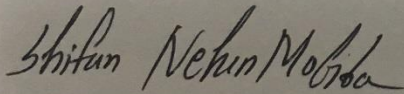
**Supervised by:**



---

**Md. Ismail Jabiullah**  
Professor  
Department of CSE  
Daffodil International University

**Submitted by:**



---

**Shifun Neher Mofida**  
ID: 151-15-4984  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Md. Ismail Jabiullah, Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Natural Language Processing*” 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.

We would like to express our heartiest gratitude to Dr. Syed Akhter Hossain, Professor and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

Finally, we must acknowledge with due respect the constant support and patients of our parents.

## ABSTRACT

In this current world where everyday people are generating a large amount of data and different business organizations are becoming more and more dependent on it, therefore it has become very important to come out of the traditional methods of data analysis and focusing on the techniques that can prepare much more accurate and valid result to make business decision making more easy and simple. This thesis proposes a technique to collect and analyze twitter posts based on different keyword based product searching to generate products market statistical report. Using this program it can be determined that if any product is getting popularity or losing its market. Few types of results were generated in this project. Each of them have their own type of importance. Overall this type of application can be a trusted support for business analyst or decision makers.

## **TABLE OF CONTENTS**

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of the Study	3
1.4 Research Questions	3
1.5 Expected Output	4
1.6 Report layout.	4
<b>CHAPTER 2: BACKGROUND</b>	<b>5-11</b>
2.1 Introduction	5
2.2 Related Works	7
2.3 Research Summery	8
2.4 Scope of the Problem	10
2.5 Challenges	10
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>12-26</b>
3.1 Introduction	12
3.2 Research Subject and Instrumentation	12
3.3 Data Collection Procedure	12

3.3.1 Creating Twitter App	13
3.3.2 R Setup	15
3.3.3 Required Library Installation	17
3.3.4 Collecting Twitter Data	18
3.4 Statistical Analysis	20
3.4.1 Data pre-processing	20
3.4.2 Data Analyze	23
3.4.3 Data Flow Model	25
3.5 Implementation Requirements	25
<b>CHAPTER 4: EXPERIMENT RESULTS AND DISCUSSION</b>	<b>27-38</b>
4.1 Introduction	27
4.2 Experimental Result	27
4.2.1 Waterfall Chart	28
4.2.2 Histogram chart	29
4.2.3 Wordcloud	33
4.4 Summery	39
<b>CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLECATION</b>	<b>40-41</b>
5.1 Summary of the Study	43
5.2 Conclusion	43
5.3 Recommendations	44
5.4 Implication for Future Study	44

**LIST OF FIGURES**  
**FIGURES**

	<b>PAGE NO</b>
Figure 3.1: Creating twitter application	13
Figure 3.2: Twitter Application for Data Collection	14
Figure 3.3: Consumer Key and Consumer Secret Key	14
Figure 3.4: Access Token and Access Token Secret	15
Figure 3.5: R Download	15



Figure 3.6: R Graphical User Interface	16
Figure 3.7: RStudio Graphical User Interface	17
Figure 3.8: Sentiment related packages	17
Figure 3.9: Setting up twitter connection	18
Figure 3.10: Data visualization library	18
Figure 3.11: Connecting program with Twitter API	18
Figure 3.12: Script for collecting twitter data based on keyword	19
Figure 3.13: Raw data collected from twitter	19
Figure 3.14: Data cleaning script	20
Figure 3.15: Data Scoring Script	21
Figure 3.16: Score for Oneplus phone	22
Figure 3.17: Score for Samsung phone	22
Figure 3.18: Score for iPhone phone	22
Figure 3.19: Score for Huawei phone	23
Figure 3.20: Score for OPPO phone	23
Figure 3.21: Negative Words	24
Figure 3.22: Positive words	24

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.23: Data Flow Diagram	25
Figure 4.1: Script for waterfall plotting	28
Figure 4.2: Waterfall Chart output	29
Figure 4.3: Histogram scoring system table	30
Figure 4.4: Scoring histogram plot / bar chart program	32
Figure 4.5: Histogram chart output	32
Figure 4.6: Word Counter Scoring	33
Figure 4.7: Wordcloud generator script	34
Figure 4.8: Wordcloud of Oneplus	35

Figure 4.9: Wordcloud of Samsung	35
Figure 4.10: Wordcloud of iPhone	36
Figure 4.11: Wordcloud of Huawei	37
Figure 4.12: Wordcloud of OPPO	37

# 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.

Supposition primarily identifies with sentiments; frames of mind, feelings and assessments. Sentiment Analysis alludes to the act of applying communication method and Text Analysis strategies to acknowledge and separate abstract info from slightly of content. Associate individual's conclusion or emotions are sometimes abstract and not realities. That intends to precisely investigate an individual's conclusion or mind-set from slightly of content are very onerous. With Sentiment Analysis from a book investigation perspective, we have a tendency to tend to are primarily hoping to induce a comprehension of the frame of mind of associate author relating to some extent throughout a touch little bit of content and its extremity; in spite of whether or not or not it's positive, negative or impartial. Recently there has been a permanent increment in enthusiasm from brands, organizations and scientists in Sentiment Analysis and its application to business investigation. The business world today, a similar because the case in varied data examination streams, are searching for Business Insight.

## 1.2 Motivation

In a centered market whereby corporations compete for customers; supporter fulfillment is viewed as a key individual. Studies state that a 'completely consummated client' produces a pair of.6 occasions additional financial gain than a 'fairly consummated client' and multiple times more income than a 'to a point discomfited client'. Feeling examination in business, otherwise referred to as sentiment mining could be a procedure of characteristic and categorization a small amount of content as per the tone passed on by it. Shopper loyalty is that the most important marker of however probably a shopper can build obtain afterward. Organizations UN agency prevail in those decrease throat things are those that build shopper loyalty a key part of their business procedure. Supposition investigation in business will demonstrate to be a major step forward for the overall whole rejuvenation.

Most social destinations give well-known objective market activity, because it is also that action is extra subject to additional insights which will be explored through a target cluster of audience interest. Audit sources are chiefly location, model Twitter, Facebook and then on. It's basic to understanding the poor, and even impartial sense qualities. The key to running a fruitful business with assessment data is that the ability to misuse unstructured measurements for vital bits of information. Consultants agree. As indicated by Bruce Temkin, a client expertise visionary, "Passionate" is one in all the 3 major expertise segments. Forrester places an infatuated commitment to the very best purpose of its client expertise pyramid.

At this era of technology sentiment analysis is a growing sector and upcoming many technology will include emotion prediction for the welfare of human being. Those areas includes business, crime control, social development and so on. Sentiment analysis is a continuous area of research. Many analysts are focusing on proposing a highly accurate sequence count to isolate the feelings of writing. Research states that grouping content at the record level or at the sentence level does not require significant detail on all parts of the element that is required in many applications. The use of the spirit of inquiry in business cannot be ignored. Opinion scrutiny in business can demonstrate a significant leap forward

for total brand rejuvenation. The way to maintain an effective business with opinion information is the ability to misuse unstructured information for notable bits of knowledge.

AI models, which largely rely on anatomically created highlights prior to characterization, have fixed this requirement for as long as hardly any years.

### **1.3 Rational of the study**

Research is a joint venture of an issue where there is an attempt to seek answers for an issue. To get the right arrangement of a right issue, clearly characterized goals are important. Simplified destinations edit the way a scientist needs to continue. Research goals are usually communicated in late terms and a lot is directed to the client in relation to the expert. The research goal may be associated with a theory or used as a mission statement in an investigation that does not involve speculation. An expert goal is a clear, compact, explanatory proclamation, which gives guidance for examining factors. For most approaches the best approach is to look at the variables in the target center, for example, to distinguish or delineate them. At some points, the goal is coordinated to distinguish the relationship between two factors. The research goal diagram gives particular objectives when the investigation is completed.

At this era of technology sentiment analysis is a growing sector and upcoming many technology will include emotion prediction for the welfare of human being. Those areas includes business, crime control, and social development and so on. The main objective of this research are:

- To use data analysis techniques to make business plan.
- To understand business trends.
- To know customer satisfaction reports about the business.
- To know product demand on the market.
- To use sentiment analysis to predict user demand for the product.
- To integrate social media with business analysis for better result.

## **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 choose “Market evaluation using lexicon based 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. 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.

## **1.6 Report Layout**

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

Chapter One: Introduction, Motivation, Rational of the Study, Research Questions, Expected Output, Report layout.

Chapter Two: Introduction, Related works, Research Summery, Scope of the Problem, challenges.

Chapter Three: Introduction, Research Subject and Instrumentation, Data Collection Procedure, Statistical Analysis, Implementation Requirements.

Chapter Four: Introduction, Experimental Result, Descriptive Analysis, Summery.

Chapter Five: Summery of the Study, Conclusion, Recommendations, Implication for Further Study.

## **CHAPTER 2 BACKGROUND**

### **2.1 Introduction**

Sentiment analysis is a computational procedure to group human feeling in various classes. Conclusion examination is workable for content, sound, video or picture. Be that as it may, every one of them will require diverse way, approach and calculation. As I took a shot at content examination, so this research is containing information about just content mining. Yet, in the event that I need to begin sentiment analysis, from the start I have to think about the notion examination and its methodologies.

Sentiment can be considered as feeling, frame of mind, or assessment. Computationally recognizing and preparing some content, we can do Sentiment investigation to decide if the author's feeling on a specific point or a subject or an item or an individual is certain, negative or nonpartisan. So utilizing regular language preparing arranging, grouping or factually seeing some systematic report from a lot of content is commonly considered as notion examination.

We 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 we are going describe those levels of sentiment analysis:

Document Level Sentiment Analysis: In Document Level Sentiment Analysis we will consider a solitary report which will have a solitary subject. Along these lines this degree of assumption examination isn't pertinent for investigation of different gatherings or blog locales. The primary test for archive level investigation might be a few messages that are not pertinent to the subject. So at whatever point we apply it, we need to evacuate unimportant sentences. For this degree of examination we can utilize both supervised and unhelpful learning. An administered learning calculation, for example, innocent portion [6] or vector machine and confused learning calculation, for example, grouping or k-implies calculation can be utilized for this arrangement.

Sentence level Sentiment Analysis: In sentence level feeling investigation we will think about each sentence in the report. So for each degree of report as general or archive or blog locales this level opinion examination is applied. Here by deciding positive and negative words we will have the option to decide if the sentence is a positive or negative sentence. Sentence level sense expectation isn't pertinent for complex sentences. Like Document Level Sentiment Prediction, we can apply both regulated and unsubsidized learning in Sentiment Prediction.

Phrase level Sentiment Analysis: In expression level feeling investigation we will consider words identified with conclusion or feeling. The degree of this forecast is a directed methodology toward expectation of sensation. Now and again the definite thought can be removed from the content on a particular point. Be that as it may, this level examination goes up against the issue of disregard and long-separation reliance. Words that show up near one another are viewed as an expression here. Sentence level investigation is applied to this examination among every one of them. Since sentence level examination can be utilized for both regulated and unhelpful learning, the utilization of both administered and inaccessible learning gives better investigation results.

Assessment expectation is principally significant for business organizations, political gatherings, and different social associations. Business organizations can decide the following methodology and accomplishment of their business by gathering client surveys



and fulfillment. Political gatherings can control existing figures and individuals utilizing the most examined points, so they will have the option to realize what they need from them. Then again, different associations will have the option to decide support for them for a specific assignment or their work. They are a few models for the significance of notion examination. So we can say that by breaking down the supposition we can profit in our handy life. The advanced time has brought an enormous field for PC researchers to work with information to decide individuals' feelings.

## **2.2 Related Works**

Sentiment Analysis is popular in light of its effectiveness. A large number of content records can be prepared in a moment or two (and for different highlights including assigned highlights, points, subjects, and so forth.), contrasted with the hours it would take a group of individuals to finish physically. Since it is so effective (and exact - Cementria has 80% precision for English substance) numerous organizations are receiving content and notion analysis and joining it into their procedures [2].

Individuals everywhere throughout the world are presently chipping away at various themes of sentiment analysis one day. As the measure of information is expanding step by step, information mining and notion investigation become increasingly well known for individuals. Different organizations require conclusion investigation to gather client data. Since it is the most effortless approach to discover their clients and target them for their business. Indeed, even now online life needs enthusiastic consistency. Utilizing conclusion examination they decide the advertising methodology, improve battle achievement, improve item informing and improve client assistance.

Many individuals around the globe are presently chipping away at foreseeing sentiment. Continuous research brought us new highlights and progressively exact outcomes. Research by Pang and Li [7] brought a wide range of approaches, for example, human articulation location; Classifying sentences as positive, negative or unbiased; Detection of

abstract and target sentences; Classifying human feelings into various classes like displeasure, joy, distress, and so on. Use of sentiment analysis in different fields [8].

Hatzivasiloglu and McKeave [9] and Esuli and Sebastiani [10] worked with extremity recognition from phrases. Yu and Hatziwasiloglu [11] and Kim and Howie [12] took a shot at as far as possible and discovered that Twitter message investigation is like sentence level estimation examination [13]. However, as of late, such huge numbers of individuals are taking a shot at different subjects of passionate expectation like Twitter, Facebook, papers, online journals, books and so forth. For instance, Safa ben Hamouda and Jalal Ekaichi [14] dealt with assumption order from Facebook for the Arabic-language time. Zhaoxia WANG, Victor Xu Chuan Tong and David Chan [15] were taking a shot at information investigation issues of online life with new calculations; Jonathan Bright, Helen Margates, Scott Hale, and Taha Yasseri [16] dealt with the utilization of web based life for inquire about. Mika Viking Mäntylä, Daniel Graziotin and Miikka Kuutila [17] took a shot at the improvement of sentiment analysis. S. Padmaja, Proc. s. Samin Fatima and Sasidhar Bandhu [18] worked in the paper on sentiment analysis and advancement of invalidation.

Each one of those scientists chipped away at subjects that depended on administered and unhelpful learning. Indeed, even sometimes cross breed techniques are material for sentiment analysis. The utilization of various calculations makes the supposition analysis process simpler. Content based sentence analysis faces an assortment of challenges. As per Professor Bing Liu from the University of Chicago, University of Computer Science, it is hard to find out precision for examination and it relies upon the degree of investigation, number of informational indexes, estimations, etc [19]. One of the most widely recognized issues is the various implications for a similar book. There are likewise different issues, for example, various dialects, alternate route words composed by the writer, composing botches, and so forth. Thusly when working with those sorts of issues it is hard to choose how to tackle them [20]. Then again, a negative word or sentence, for example, 'This isn't great' can mess up calculations to discover the precise outcome, to decide if it is a positive or negative word, a positive word and a negative word appearing as positive.

### 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 analysis is a subject of incredible intrigue and improvement as it has many viable applications. As freely and secretly accessible data on the Internet is ever growing, countless communication sentiments are accessible in survey sites, discussions, sites, and web-based life. This unstructured data can naturally be transformed into organized information of general assessments about items, administrations, brands, government issues, or any topic about which individuals have feelings, with the help of realizing the framework of the forecast. Can express This information can be exceptionally valuable for business applications such as advertising testing, advertising, item audit, net advertiser scoring, item input, and client support.

In the particular correspondence area, Sentiment Prediction are a rapidly evolving topic. With the expansion into Internet-based life, online retail, and personal web journals and presentations where there is a tinge of openness to know, the conclusion has turned into a quick development in expectation that can turn into a key competency.

At the point when we expect a sentiment on a substance, the principle that sees through the point is the emotions in the substance and we are raising the skew dependent on those conclusions. A combination is an assertiveness consisting of two major parts. One of them is an objective or point and the other is an assessment on the subject. Consider a sentence, "I love this organization", here "this organization" is the point and the action is a certain notion of communicating with the word "love".

Sentiment analysis is not just an element in a social inquiry tool - it is an area of study. This area is yet to be investigated, yet is not of extraordinary length due to the versatile design of this investigation, similarly parts of the semantics are still to be discussed or not fully understood. Current methods of dealing with perception examination can be gathered into three primary classifications: information-based strategy, measurable techniques, and cross breed pass.

Information-based systems impact the message of unobservable classes that rely on the dependence of words with indeterminate effects, for example, excited, sad, apprehensive, and tedious. Some learning bases make a list of words with obvious effects, yet discretionary words are a potential "bias" for specific emotions.

Cross-breed influences both AI and components from airports and learning portraits, for example, cosmology and semantic arrangements, so that semantics can be communicated in a different way, for example, through examination of ideas that Implicit data are not disproportionately passed on, however, which are associated with different views, which it does in this way.

Open source programming apparets create AI, Insights and specialized languages, including a computerized strategy of guessing at the vast accumulation of writing, including pages, online news, web exchange gathering, online audits, web journals, and web-based social networking then, using openly accessible assets, separating to feel semantic and data related to different language views. Estimation can likewise be predicted on visual matter, ie pictures and recordings. One of the main approaches to this path is SENTIBANK using a modifier thing pair depiction of visual content [1].

Applications for perception forecasting are permanent. More and more we are using it to follow online life checks and client audits in VOCs, study responses, contenders, and the like. Be that as it may, it is additional down to earth for use in professional examination and in situations in which the material must be broken.

## **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 an understanding for the machine. Therefore people are working hard on the algorithm, sentiment word library and different techniques. In this research we are using rule based technique known as lexicon based analysis. This research focuses on the best and accurate result generation using the algorithm.

On the other hand another major challenge of this research is finding whether the text means any actual sentiment related class or not. Even some cases actual sentiment call cannot be found because of use of negation words.

## **2.5 Challenges**

One of the significant issues of sentiment analysis is precision now. Numerous specialized or theoretical provokes become an impediment to breaking down the accurate importance of feeling and recognizing suitable feeling extremity. Sentiment analysis is the act of applying normal language handling and content examination methods to recognize and extricate passionate data from the content [3]. The level of exactness issue is hard to reply, said Bing Liu, a Chicago software engineering educator having some expertise in information mining. It relies upon what the estimation will be, the degree of investigating the content, and the quantity of informational collections in the space and the sound nature of the video, among different factors. By the by, he imagines that progress is being made in such manner [2].

It is all the more testing to investigate feeling/feeling all the more profoundly in feeling forecast. Positive and negative is a very straightforward examination however testing is to take out feelings, for example, how much detest is there inside the supposition, how much delight, how much pity, and so on. 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 project I tried to analyze twitter data to get sentimental state from it. To do that at first I need to collect data from twitter. I created a twitter app to get the data. Then I connected it with my application. I wrote a script to analyze collected data. So when I run the program it will automatically collect the data, analyze it and give the output.

## **3.2 Research Subject and Instrumentation**

As I selected Market evaluation using lexicon based sentiment analysis 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 analysis.

I found 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 selected 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 I started learning R programming language which is mainly used for data analysis and it is a high level programming language

After that I created a twitter app for data collection and using R language I 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.

### **3.3.1 Creating Twitter App**

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

## Create an application

**Application Details**

**Name \***

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

**Description \***

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

**Website \***

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)

**Callback URLs**

Where should we return after successfully authenticating? OAuth 1.0a applications must explicitly specify their oauth\_callback URL(s) here, as well as include the one of the URLs below in the request token step. To restrict your application from using callbacks, leave this field blank.

**Developer Agreement**

Yes, I have read and agree to the Twitter Developer Agreement.

Figure 3.1: Creating twitter application

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



## Twitter Apps

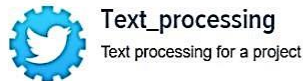
[Create New App](#)

Figure 3.2: Twitter Application for Data Collection

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 those information no one will be able to connect this app with my application.

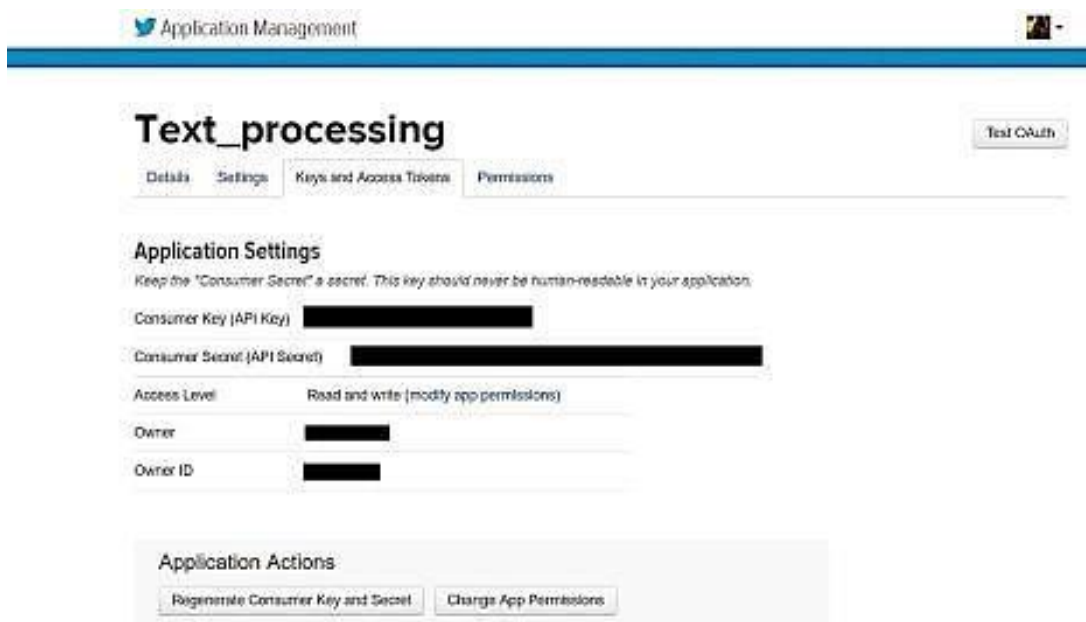


Figure 3.3: Consumer Key and Consumer Secret Key

## Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	[REDACTED]
Access Token Secret	[REDACTED]
Access Level	Read and write
Owner	[REDACTED]
Owner ID	[REDACTED]

Figure 3.4: Access Token and Access Token Secret

Those codes will be used when the system will try to scrap data from twitter for data analysis. It must be considered that leak of those access key may harm someone's twitter account.

### 3.3.2 R Setup

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

R-3.3.2 for Windows (32/64 bit)

[Download R 3.3.2 for Windows](#) (62 megabytes, 32/64 bit)  
[Installation and other instructions](#)  
[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is [CRAN\\_MIRROR> bin/windows/base/release.htm](#).

---

Last change: 2016-10-31, by Duncan Murdoch

Figure 3.5: R Download

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

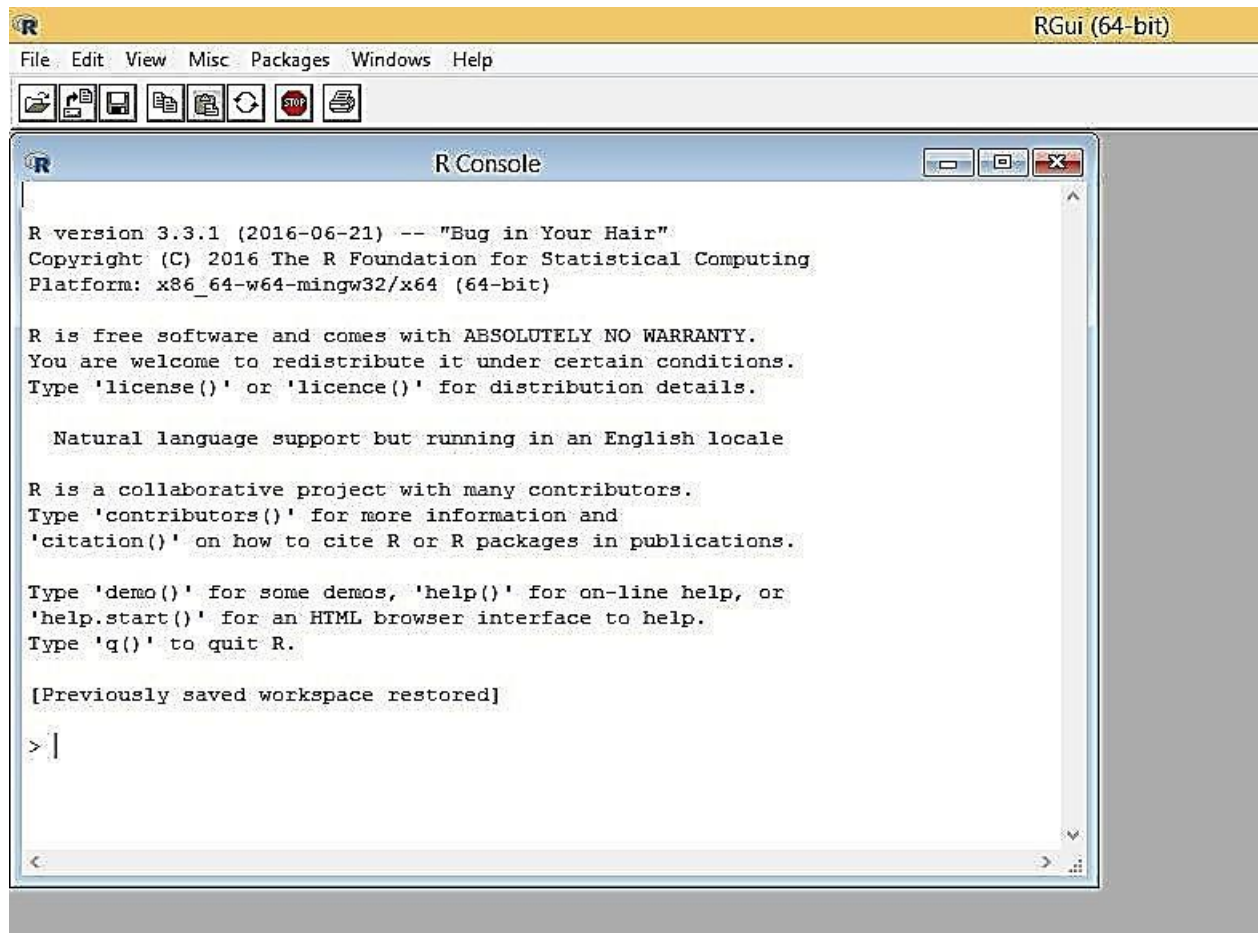


Figure 3.6: R Graphical User Interface

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 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.

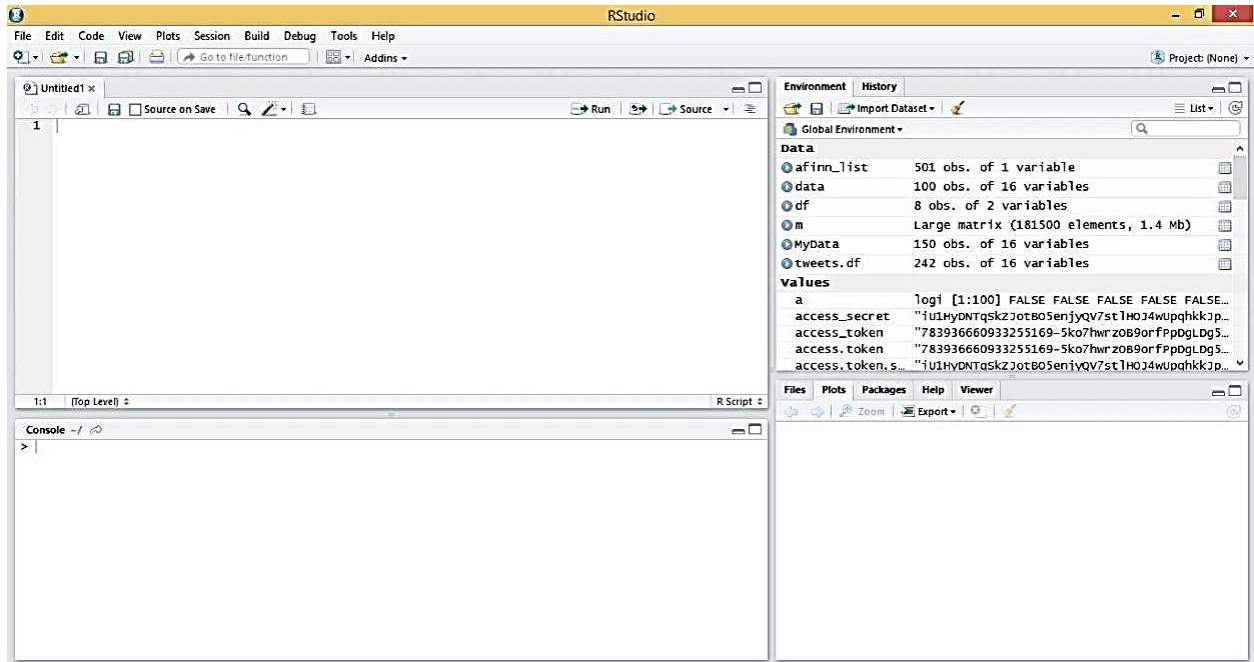


Figure 3.7: RStudio Graphical User Interface

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

### 3.3.3 Required Library Installation

After setting up all the necessary components for the application it is required to install R library files needed to analyze sentiment and connect twitter with the system.

For sentiment analysis required packages are “sentimentr”, “stringr” and “plyr”. Those will be installed as given below-

```
#Installing the sentimentr package
#install.packages("sentimentr")
library("sentimentr")
library("stringr")
library("plyr")
```

Figure 3.8: Sentiment related packages

For connecting this program with twitter required packages are “tweetR”, “httr”, “tm” and “SnowballC”. Those need to be installed as given below-

```

#Setting up Twitter:

#Install the required packages
#install.packages("twitterR")

#twitterR acts as an interface between R and Twitter and helps to scrap Twitter data
library("twitterR")
library("httr")
library("tm")
library("snowballc")

```

Figure 3.9: Setting up twitter connection

After analysis data output will be visualized by those packages - “lattice”, “wordcloud” and “RColorBrewer”. Packages need to be loaded by given way below-

```

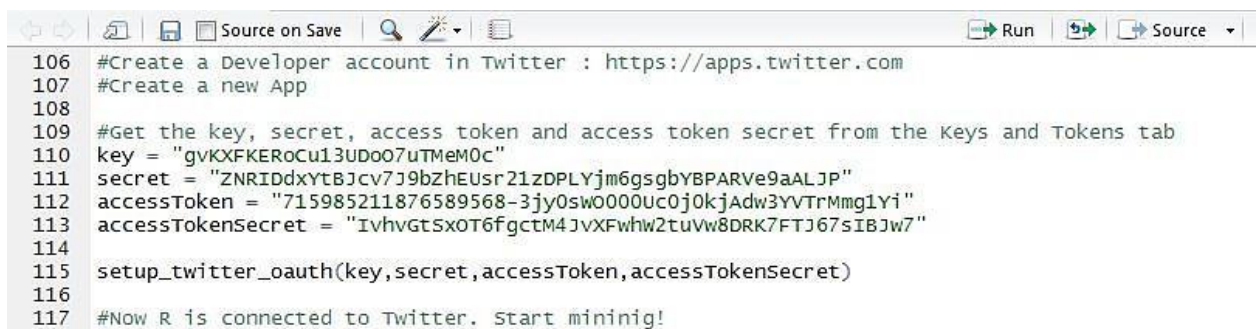
#Importing the 'lattice' package for data visualization
library("lattice")
#Importing the 'wordcloud' & 'RColorBrewer' packages
library("wordcloud")
library("RColorBrewer")

```

Figure 3.10: Data visualization library

### 3.3.4 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:



```

106 #Create a Developer account in Twitter : https://apps.twitter.com
107 #Create a new App
108
109 #Get the key, secret, access token and access token secret from the keys and Tokens tab
110 key = "gvKXFKERoCu13UDo07uTMeM0c"
111 secret = "ZNRIDdxYtBJcv7J9bZHEusr21zDPLYjm6gsgbYBPARVe9aALJP"
112 accessToken = "715985211876589568-3jy0sw0000uc0j0kjAdw3YVTrMmg1Yi"
113 accessTokenSecret = "IvhvGtSxOT6fgctM4JvXFwhw2tuVw8DRK7FTJ67sIBJw7"
114
115 setup_twitter_oauth(key,secret,accessToken,accessTokenSecret)
116
117 #Now R is connected to Twitter. Start mininig!

```

Figure 3.11: Connecting program with Twitter API

```

#Compare tweets of phones:
#Get the tweets of "ONEPLUS", "SAMSUNG", "IPHONE", "HUAWEI" and "OPPO"
#optweets = searchTwitter("anykeyword", n=1000, lang="en", since = "2017-12-10", until = "201
optweets = searchTwitter("oneplus", n=1000, lang="en")
satweets = searchTwitter("samsung", n=1000, lang="en")
iptweets = searchTwitter("iphone", n=1000, lang="en")
hutweets = searchTwitter("huawei", n=1000, lang="en")
optweets = searchTwitter("oppo", n=1000, lang="en")

#Get only texts from the tweets
op_txt = sapply(optweets, function(x) x$text())
sa_txt = sapply(satweets, function(x) x$text())
ip_txt = sapply(iptweets, function(x) x$text())
hu_txt = sapply(hutweets, function(x) x$text())
op_txt = sapply(optweets, function(x) x$text())

```

Figure 3.12: Script for collecting twitter data based on keyword

After collecting tweet using the program previously discussed I will get raw tweet as given below-

	text
1	@arbuzzzy Harris was our best reliever the second ha...
2	@oppo_researcher According to Google, this is Alask...
3	trying to win a trip to new delhi because why not? #Co...
4	#Spark, #OPPO #NewZealand strengthen partnershi...
5	@Ritiya Hi Ritesh, we would like to inform you that OP...
6	@oppo_researcher I have a MPA. I've handed Rob Por...
7	RT @phone_stores: Pls Take a sec To RT<ed><U+00A0...
8	RT @colorsglobal: Giveaway for a trip to New Delhi! F...
9	RT @colorsglobal: Giveaway for a trip to New Delhi! F...
10	RT @troonaccies: Cracking win for us this morning ru...
11	Helping a friend basig naay need phone diri for sale: ...
12	Great Selfie Camera <U+2705> Super Fast Charging ...
13	RT @colorsglobal: Giveaway for a trip to New Delhi! F...
14	RT @SuperSaf: 0-100% in under 30 mins <ed><U+00A0...
15	RT @HamiltonKal: @mebcaux @brithume Fusion GPS...

Figure 3.13: Raw data collected from twitter

I used libraries 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**

Now I have my required data for the sentiment analysis of product review. But before that data preprocessing is required. In preprocessing step unexpected data columns and other unnecessary things will be removed to make data clear and easy to analyze. Because of that this can be divided in two sectors- Data pre-processing and Data analysis. After completing those steps I will be able to get my expected output.

#### **3.4.1 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-

```

#Removing punctuation using global substitute
sentence = gsub("[[:punct:]]", "", sentence)
#Removing control characters
sentence = gsub("[[:cntrl:]]", "", sentence)
#Removing digits
sentence = gsub('\\d+', '', sentence)

#Error handling while trying to get the text in lower case
tryTolower = function(x)
{
  #Create missing value
  y = NA
  #Try Catch error
  try_error = tryCatch(tolower(x), error=function(e) e)
  #If not an error
  if (!inherits(try_error, "error"))
    y = tolower(x)
  #Return the result
  return(y)
}

```

Figure 3.14: Data cleaning script

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

After preprocessing the data will be clean and accordingly scoring function will score the tweet based on lexicon based sentiment analysis.

Script for scoring the tweet is given below-



```

#Use this tryTolower function in sapply
sentence = sapply(sentence, tryTolower)

#split sentences into words with str_split (stringr package)
word.list = str_split(sentence, "\\s+")
#Unlist produces a vector which contains all atomic components in word.list
words = unlist(word.list)

#Compare these words to the dictionaries of positive & negative words
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)
#Example: If a sentence is "He is a good boy",
#then, pos.matches returns: [NA, NA, NA, *some number*, NA] : the number depends on the dictionary
#neg.matches returns: [NA, NA, NA, NA, NA]
#So the output has NA's and numbers
#We just want a TRUE/FALSE value for the pos.matches & neg.matches
#Getting the position of the matched term or NA
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)
#This would return : [F, F, F, T, F] depending on the NA or the match
#The TRUE or FALSE values are treated as 1/0
#To get the final score of the sentence:
score = sum(pos.matches) - sum(neg.matches)
return(score)
}, pos.words, neg.words, .progress=.progress )

#Now the scores are put in a dataframe and returned
scores.df = data.frame(text=sentences, score=scores)
return(scores.df)
}

```

Figure 3.15: Data Scoring Script

Finally the preprocessing and scoring will return tweet score for all the phone brands. The sample result is given below-

	text	score	phone	very.pos	very.neg
1	@arbuzzzy Harris was our best reliever the second ha...	1	ONEPLUS	0	0
2	@oppo_researcher According to Google, this is Alask...	1	ONEPLUS	0	0
3	trying to win a trip to new delhi because why not? #Co...	1	ONEPLUS	0	0
4	#Spark, #OPPO #NewZealand strengthen partnershi...	0	ONEPLUS	0	0
5	@Ritiya Hi Ritesh, we would like to inform you that OP...	2	ONEPLUS	1	0
6	@oppo_researcher I have a MPA. I've handed Rob Por...	0	ONEPLUS	0	0
7	RT @phone_stores: Pls Take a sec To RT<ed><U+00A0...>	2	ONEPLUS	1	0
8	RT @colorosglobal: Giveaway for a trip to New Delhi! F...	2	ONEPLUS	1	0
9	RT @colorosglobal: Giveaway for a trip to New Delhi! F...	2	ONEPLUS	1	0
10	RT @traonaccies: Cracking win for us this morning ru...	3	ONEPLUS	1	0

Figure 3.16: Score for Oneplus phone

	text	score	phone	very.pos	very.neg
1000	RT @oppo: #OPPOReno2, our new quad camera smart...	0	ONEPLUS	0	0
1001	If anyone comes across articles with Samsung's pers...	1	SAMSUNG	0	0
1002	Social perfectionism, labour markets & privilege...	2	SAMSUNG	1	0
1003	@loomnetwork @neondistrictRPC @LevX @AxieInfini...	0	SAMSUNG	0	0
1004	RT @SamsungMobile: Art to honor the dearly departe...	1	SAMSUNG	0	0
1005	RT @MayorTswiit: People are jus gbas gbos themselv...	0	SAMSUNG	0	0
1006	Springboks Win the Rugby World Cup and Hearts of t...	1	SAMSUNG	0	0
1007	6 pcs Eyeshadow Makeup Brushes IKSMarkets - Thing...	0	SAMSUNG	0	0
1008	Now up on the auction block: Samsung Galaxy S6 32...	1	SAMSUNG	0	0
1009	@SpotifyCares I forgot both my password and usern...	0	SAMSUNG	0	0
1010	@Huawei But probably 10 times the amount for sam...	0	SAMSUNG	0	0
1011	RT @MikeFeibus: Wow, lots to say about @Google buyi...	1	SAMSUNG	0	0

Figure 3.17: Score for Samsung phone

	text	score	phone	very.pos	very.neg
2001	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2002	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2003	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2004	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2005	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2006	@ummachikutti iPhone 6s	0	IPHONE	0	0
2007	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2008	RT @smnthagcaoli: @_bellughhh sana all maganda,...	1	IPHONE	0	0
2009	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2010	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0
2011	RT @RelaxedReward: HUGE GIVEAWAY! <ed><U+00A0...	0	IPHONE	0	0

Figure 3.18: Score for iPhone phone

	text	score	phone	very.pos	very.neg
3003	RT @Huawei: When you work, fight and spend so muc...	2	HUAWEI	1	0
3004	Swiss Telecom CEO Explains Why He's Sticking With ...	0	HUAWEI	0	0
3005	@mg0314a "At Huawei were are on the cutting edge ...	0	HUAWEI	0	0
3006	@vtchakarova Russia and Huawei team up as tech c...	-1	HUAWEI	0	0
3007	RT @CBK1320: POTENTIALLY GREAT NEWS, Folks Less ...	2	HUAWEI	1	0
3008	RT @TADHackJHB: Congratulations to @YolandaMab...	4	HUAWEI	1	0
3009	@Huawei But probably 10 times the amount for sam...	0	HUAWEI	0	0
3010	RT @iam_jobaba: Pronounce Huawei and you'll get th...	0	HUAWEI	0	0
3011	RT @patrickbetdavid: Very creative ad by @Huawei i...	3	HUAWEI	1	0
3012	@mike_pence @realDonaldTrump @MattBevin "The ...	-1	HUAWEI	0	0
3013	RT @huawei_global: Here's how to get your free year ...	1	HUAWEI	0	0

Figure 3.19: Score for Huawei phone

	text	score	phone	very.pos	very.neg
4004	#Spark, #OPPO #NewZealand strengthen partnershi...	0	OPPO	0	0
4005	@Ritiya Hi Ritesh, we would like to inform you that OP...	2	OPPO	1	0
4006	@oppo_researcher I have a MPA. I've handed Rob Por...	0	OPPO	0	0
4007	RT @phone_stores: Pls Take a sec To RT<ed><U+00A0...	2	OPPO	1	0
4008	RT @colorosglobal: Giveaway for a trip to New Delhi! F...	2	OPPO	1	0
4009	RT @colorosglobal: Giveaway for a trip to New Delhi! F...	2	OPPO	1	0
4010	RT @troonaccies: Cracking win for us this morning ru...	3	OPPO	1	0
4011	Helping a friend basig naay need phone diri for sale: ...	1	OPPO	0	0
4012	Great Selfie Camera <U+2705> Super Fast Charging ...	3	OPPO	1	0
4013	RT @colorosglobal: Giveaway for a trip to New Delhi! F...	2	OPPO	1	0
4014	RT @SuperSaf: 0-100% in under 30 mins <ed><U+00A...	1	OPPO	0	0
4015	RT @HamiltonKal: @mebcaux @brithume Fusion GPS...	0	OPPO	0	0

Figure 3.20: Score for OPPO phone

### 3.4.2 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 Classifier by using lexical analysis structure, Combination of lexicon and as well as learning based approach that is used for

concept level analysis, interdependent latent dirichlet allocation, combined model of feature extraction and opinion miner. In this project I used Lexicon based approaches for sentence level sentiment analysis. Here I first divided my tweet sentences into words. So that I am 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:

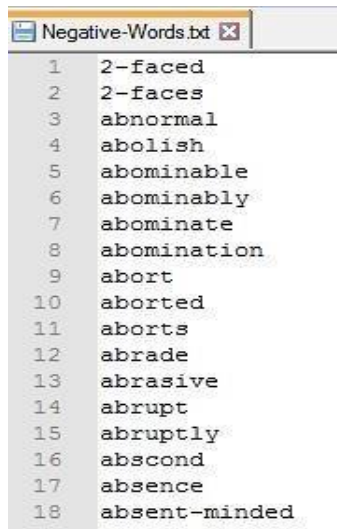


Figure 3.21: Negative Words

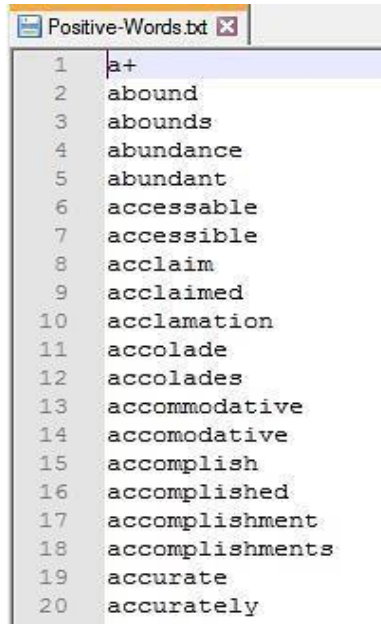


Figure 3.22: Positive words

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

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

### 3.4.3 Data Flow Model

In this project I divided my data flow in 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 connects it with R, so now data collection is ready. When data collection is complete then R script analyzes it and gives some statistical report as bar diagram, word cloud as output. Then all the outputs are sent to the browser and browser visualizes them for user.

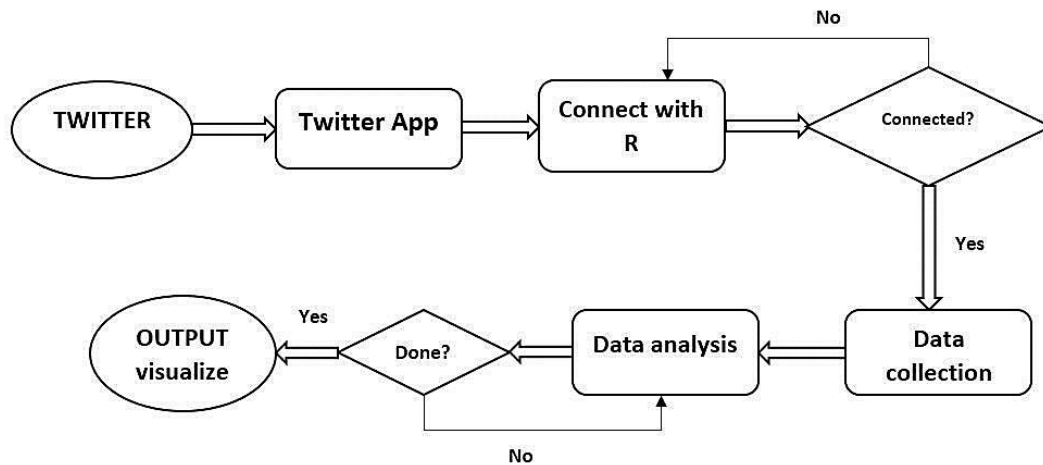


Figure 3.23: Data Flow Diagram

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 3.25.

### 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 we just follow the instruction as I discussed before about creating application in twitter we will get some access token keys. Now my twitter part is ready. After that R software need to be installed and based on what we are going to do in the project required libraries needed to be installed in R platform. Than the system is ready to implement the project.

So if I list all those things, it will be –

1. Twitter API (Developer Application)
2. R
3. RStudio
4. Installed Libraries (required)

## 5. Internet connection

# **CHAPTER 4 EXPERIMENT RESULTS AND DISCUSSION**

## **4.1 Introduction**

To use an application creating a user interface is the best option. On the other hand for a business analysis or application related to this may be much more efficient with the help of a user friendly GUI. Because this GUI will make things easy for non-technical people, even for making reports of the analysis.

## 4.2 Experimental Result

The analysis is done by R programming language with the help of twitter API. The basic connection method connects twitter with the program so that the program can collect necessary data from there. The total experimental result is divided into three types of outputs. They are-

1. Waterfall Chart
2. Histogram Chart
3. Word cloud.

The waterfall and bar chart will have the results for all the phone models in one figure. But considering the bar chart, though the results are in one figure but they all are independent. For the wordcloud section every phone model will have its own wordcloud that will have the most frequent and most common words that appeared in the tweet.

Those analysis report will show us the complete state of the sentiment for my collected tweets, so that the review report can be generated using it. 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 analyze the review of iPhone mobile 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. So after comparing few months report it will be easy to predict the growth of the business or user satisfaction from the analysis report and even more graphs or progress bars can be generated using those report.

### 4.2.1 Waterfall Chart

In this part of the analysis waterfall predicts the positivity and negativity of the phone brands. In this part positivity or negativity depends on the score calculates from the tweets by sentiment analysis. The higher position a bar goes the higher the chart has positivity. For negativity the bar will flow toward surface from the center of the chart.

The chart represents all the necessary classes of phone brand at a time so that it become easy to find the most famous one from the list. As the analysis is based on keyword that is



used in the search module. Therefore the input value of the phone brand or any kind of product can be used in this analysis. At the same time this application can be used to know about the market.

To predict the chart a simple R program is used. The script is given below-

```
#Counting the number of tweets for each phone
count = c(length(op_txt), length(sa_txt), length(ip_txt), length(hu_txt), length(op_txt))

#Joining texts
phone = c(op_txt, sa_txt, ip_txt, hu_txt, op_txt)

#Use the score.sentiment function
scores = score.sentiment(phone, pos, neg, .progress='text')

#Adding variables to 'scores' dataframe
scores$phone = factor(rep(c("ONEPLUS", "SAMSUNG", "IPHONE", "HUAWEI", "OPPO"), count))

#Calculating the number of very positive and very negative tweets
scores$very.pos = as.numeric(scores$score >= 2)
scores$very.neg = as.numeric(scores$score <= -2)
numpos = sum(scores$very.pos)
numneg = sum(scores$very.neg)

#Calculating the global sentiment score
global_score = round( 100 * numpos / (numpos + numneg) )
head(scores)

#Plotting a Box Plot of these phones
boxplot(score~phone, data=scores, col = c("red", "blue"))
```

Figure 4.1: Script for waterfall plotting

In this section analysis works on rule base. This analysis only consider all those scores that have at least either positive or negative 2 score. Those are defined as strong positive and strong negative. Then the number of positive and number of negative tweet is calculated.

The formula for processing this part of analysis is given below-  

$$\text{Score} = \text{number\_of\_round}(100 * \text{positive\_score} / (\text{positivr\_score} + \text{negative\_score})) \dots\dots (1)$$

Finally from the following equation the final outcome is delivered by the program. Sample output of the analysis is given below:

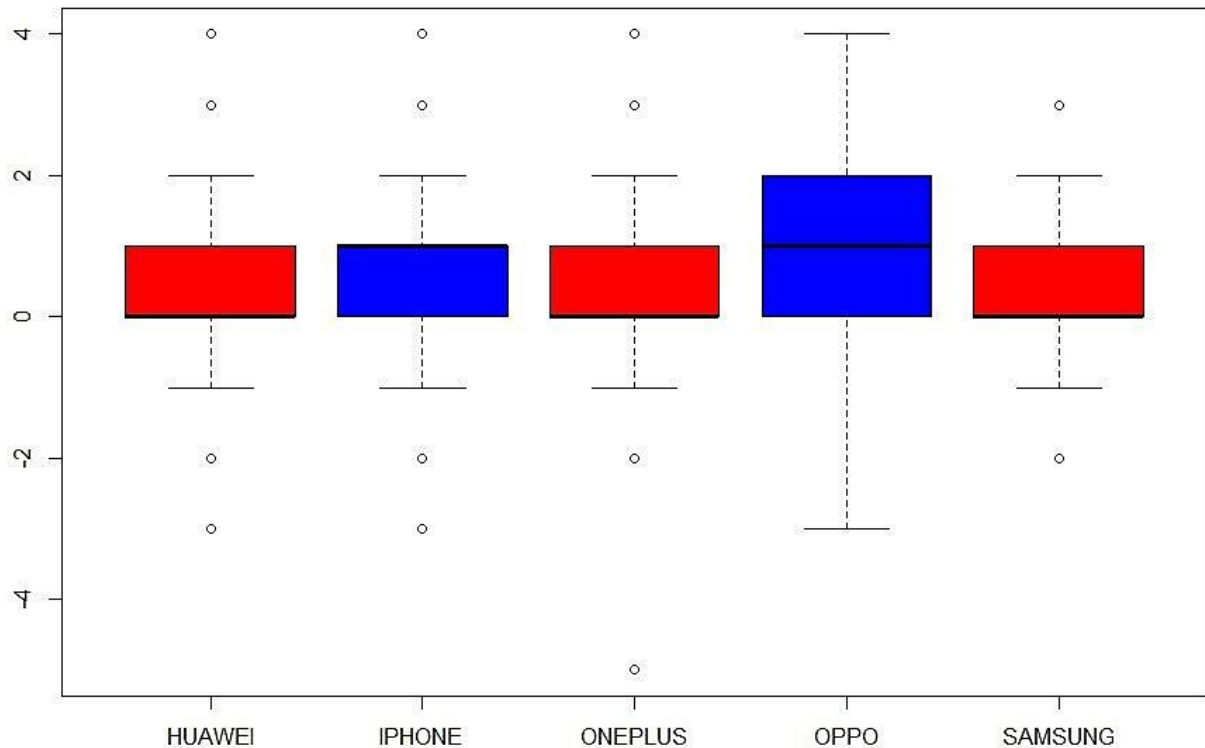


Figure 4.2: Waterfall Chart output

It can be found from the following chart that though all the phone brands have a positive impression in the market and to their customers. But among them OPPO has a superior position based on the current tweeter posts. May be it is because of their upcoming new smart devices and their features.

#### 4.2.2 Histogram Chart

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 is 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//.....](http://.....)
5. He is a good man.

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

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

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

Figure 4.3: Histogram scoring system table

For polarity classification sentiment analysis lexicon based analysis is used with the help of bag of word. So this time the application utilized inherent library for the positive and negative writings. At the point when client run the program, the program coordinates each and every word with the positive and negative words. On the off chance that match discovered, it include point either positive side or in negative side. It monitors this estimation for the last assurance of extremity. This procedure is otherwise called "Back of words".

A passage contains different sentences and in view of that the inspiration or cynicism depends of the score of the sentence. In the event that the sentence speak to positive score,

at that point the likelihood of turning out to be certain passage will be high. So I can write this as-

$$SS = PW - NW \dots\dots\dots (2)$$

From equation (1), SS is Sentence score, PW is positive word score and NW is negative word score. Each positive and negative word will get 1 point and the total score will be some of positive words and negative words.

Here in equation (1), if the score of SS is greater than 0, SS=PSS and if the score of SS is smaller than 0, SS=NSS.

Here in equation (1), if the score of SS is greater than 0, SS=PS and if the score of SS is smaller than 0, SS=NSS .

From the collected score from all the sentences of the paragraph for a single paragraph I can write the equation as:

$$SP = PSS + NSS \dots\dots\dots (3)$$

Here in equation (2), SP is the score of the paragraph. If the score of SP is positive then the paragraph expressing positive sentiment. On the other hand if SP is negative then the paragraph expressing negative sentiment.

This almost follows the same technique used in the polarity class but the basic difference is in polarity class it classify the positive, negative and neutral emotion and in this case the classifier divide the tweet based on the score of the tweet. So that the strength of the tweet can be found through this analysis.

In the following figure for polarity class user will be able to see that the application marked positive tweets bar diagram using green color, negative tweets bar diagram using red color and neutral bar is sky blue. On the top it is showing the search keyword.

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

```

#Importing the 'lattice' package for data visualization
library("lattice")

#Plotting a histogram
histogram(data=scores, ~score|phone, main="Sentiment Analysis of 5 type of Phones", col = c("re

#writing the word cloud function
constructwordcloud = function(tweettext){

#Converting them to UTF-8
tweettext=iconv(tweettext, to= "utf-8", sub="")
#Putting them in a Corpus
mycorpus = Corpus(VectorSource(tweettext))
#Data Cleaning processes
mycorpus <- tm_map(mycorpus, content_transformer(tolower))
mycorpus <- tm_map(mycorpus, removePunctuation)
mycorpus <- tm_map(mycorpus, removewords, stopwords("english"))
mycorpus <- tm_map(mycorpus, removeNumbers)

#Converting them to a DocumentTermMatrix & TermDocumentMatrix
dtm=DocumentTermMatrix(mycorpus)
tdm=TermDocumentMatrix(mycorpus)

dtmMatrix=as.matrix(dtm)
dtmMatrix

tdm2=as.matrix(tdm)
tdm2

#Getting the frequency of the words
frequency=colSums(dtmMatrix)
frequency=sort(frequency,decreasing = TRUE)

```

Figure 4.4: Scoring histogram plot / bar chart program

In this part, based on the analysis 5 different histogram chart was generated for each of the phone brand. Each one of them have different range of tweet score. Values on the right side represents the positive scores and scored on the left side represents negative scores. Those histogram bars were classified by the score each tweet have. For the class 1 it means all the tweet having score 1 belongs to that class. The sample output is given below-

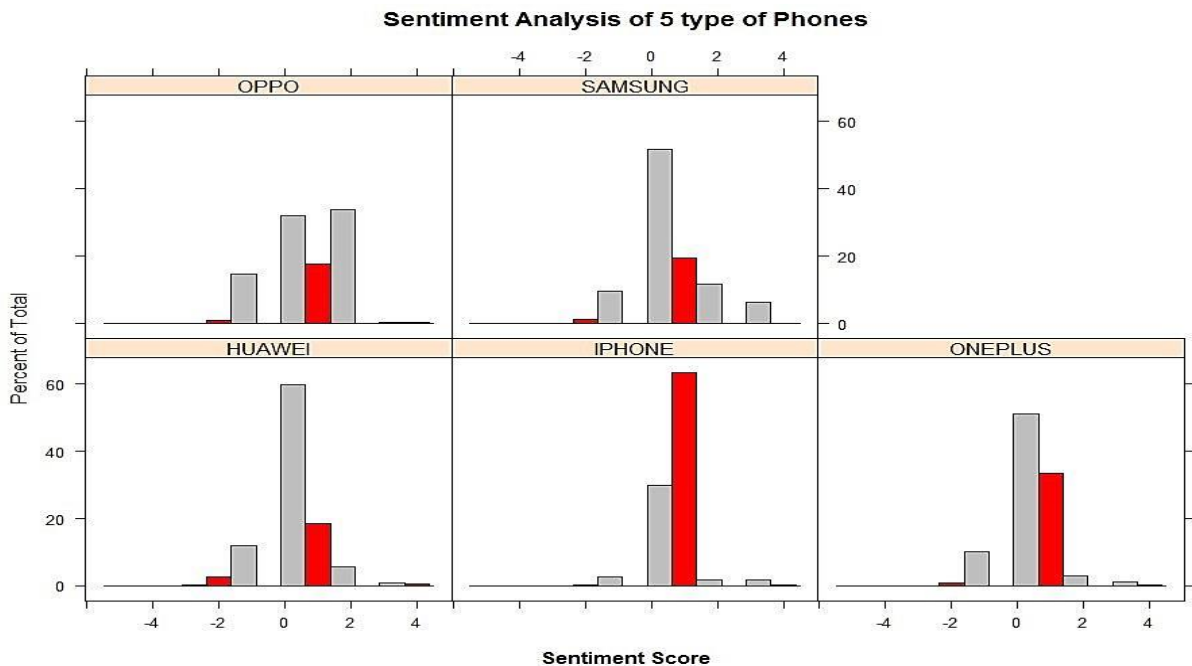


Figure 4.5: Histogram chart output

From the given figure, it can be found that in every phone brand except iPhone contains neutral or 0 score. Considering overall result, iPhone have the highest positive sentiment score. That means user of iPhone are mostly satisfied. Huawei and OPPO has a very good competition among them for scoring highest negative score. Customer satisfaction of Oneplus user is also good enough. There is a balance between positive and negative of Samsung mobile.

### **4.2.3 Wordcloud**

Wordcloud can be described as cloud of words where the algorithm finds the most common and mostly appeared words from a given data set. During the analysis the sentences are divided into words and they are tokenized so that they can be counted by the program. According to that program creates a list of words and arrange them in an order that can make it possible to plot them in wordcloud.

The uses a counter to count the same word and during word segmentation the program arranges the words in a simpler form so that the processing become easier. Counter starts from 0 and increases with each entry and it starts from the very beginning of the document and finishes at the end. If any interruption occur the program will show an error message. The sample output of word counter is given below-

	term	occurrences
<b>manag</b>	<i>manag</i>	222
<b>work</b>	<i>work</i>	202
<b>system</b>	<i>system</i>	193
<b>project</b>	<i>project</i>	184
<b>problem</b>	<i>problem</i>	171
<b>will</b>	<i>will</i>	168
<b>exampl</b>	<i>exampl</i>	164
<b>differ</b>	<i>differ</i>	157
<b>approach</b>	<i>approach</i>	154
<b>make</b>	<i>make</i>	153
<b>question</b>	<i>question</i>	152
<b>data</b>	<i>data</i>	141
<b>peopl</b>	<i>peopl</i>	141
<b>process</b>	<i>process</i>	140
<b>point</b>	<i>point</i>	135
<b>chang</b>	<i>chang</i>	130

Figure 4.6: Word Counter Scoring

Those words were counted from all the collected tweets, which means it is not the result of only one mobile brand but for all the mobile brands together.

In this step wordcloud will be generated for each of the mobile brand individually. As in cloud tiny pieces of water creates a drop, like that in a word cloud all together they mean something understandable. Not always clear meaning found from there, but a basic idea can be found.

A script for wordcloud generate this output from the data collected from twitter. The program not only considers the words that have most occurrence but also relevant. The script is given below –

```

#Importing the 'wordcloud' & 'RColorBrewer' packages
library("wordcloud")
library("RColorBrewer")

#Get the names of the words
words=names(frequency)
#Basic word cloud
wordcloud(words[1:100], frequency[1:100])

#Advanced word cloud
col <- brewer.pal(5,"Dark2")
wordcloud(words[1:100], frequency[1:100], scale = c(4,1), rot.per = 0, colors= col,random.co
}

#Constructing word clouds for the phones
constructwordcloud(sa_txt)
constructwordcloud(op_txt)
constructwordcloud(ip_txt)
constructwordcloud(hu_txt)
constructwordcloud(op_txt)

```

Figure 4.7: Wordcloud generator script

The words in a wordcloud appear in a systematic way. The words in a wordcloud are not always the same. Even they are not the same color. Their size and color depends on the score and importance in the text data. So according to the analysis, search keywords are green and most frequent words appear as purple or pink. A little lower frequent words are orange and general words that appeared many times visualize little blueish green color and they are the smallest ones.

Sample wordcloud output for Oneplus is given below:









Figure 4.10: Wordcloud of iPhone

Some common words in this wordcloud are like, giveaway, app, Trump, win, etc. So prediction can be like this, that there is an offer going on about giving away iPhones and someone can win that. On the other hand, people may talking about how much they like the new iPhone and may be Trump is going to set some new business policy.

Sample wordcloud output for Huawei is given below:



Figure 4.11: Wordcloud of Huawei

Again from this wordcloud the prediction looks like Huawei and its user are talking about the smartphone market and due to recent restrictions about Huawei the talks about the smartphone market increased.

Sample wordcloud output for OPPO is given below:



Figure 4.12: Wordcloud of OPPO

It looks like something is going through the smartphone market and something new is going to be introduced to the smartphone users. Because like Apple and Huawei, OPPO keyword also finds keywords related to market. It is true that recent smartphone technology is running too fast and as all knows that 5G technology is going to be introduced soon, so due to that the smartphone market is talking about major phone brands a lot.

#### **4.4 Summery**

Finally, it can be said that this type of analysis is capable to find market trends and predict people's sentiment regarding the product. Because, as the use of social media is increasing rapidly, people are becoming more and more dependent to it. It is very common that people post their emotions, opinions and judgement over social media. Therefore it can be a very strong source of finding future trend of the market. Therefore, according to this research my final prediction will be using sentiment analysis finding market trend and business development is quite effective and this field has a great future which will change people's life.

## **CHAPTER 5 SUMMARY, CONCLUSION, RECOMMENDATION AND IMPELICATION**

### **5.1 Summary of the Study**

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

### **5.2 Conclusion**

In this project I tried to develop a complete project using sentiment analysis of tweet data to analyze product review of user to develop business. 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 everyone will have real time data analysis experience.

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 about 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 Recommendations**

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.

### **5.4 Implication for Future Study**

This 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 this research work. In future I am 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.

## REFERENCE

- [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] Sathya, 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-formachinelearning>."
- [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., Sebastiani, 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., Hovy, 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.](https://www.quora.com/What-are-the-applications-of-sentiment-prediction)" Last Accessed: 25<sup>th</sup> August 2019.
- [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).



8%

SIMILARITY INDEX

2%

INTERNET SOURCES

8%

PUBLICATIONS

%

STUDENT PAPERS

---

PRIMARY SOURCES

---

S. M. Mazharul Hoque Chowdhury, Priyanka **1** Ghosh, Sheikh Abujar, Most. Arina Afrin, Syed Akhter Hossain. "Chapter 1 Sentiment Analysis of Tweet Data: The Study of Sentimental State of Human from Tweet Text", Springer Science and Business Media LLC, 2019

Publication

4%

**2**

S. M. Mazharul Hoque Chowdhury, Sheikh Abujar, Mohd. Saifuzzaman, Priyanka Ghosh, Syed Akhter Hossain. "Chapter 38 Sentiment Prediction Based on Lexical Analysis Using Deep Learning", Springer Science and Business Media LLC, 2019

Publication

1%

---

---

3

[towardsdatascience.com](https://towardsdatascience.com)

Internet Source

1%

---

4

"Detecting Fraud Apps using Sentiment

Research", International Journal of Recent  
Technology and Engineering, 2019

Publication

<1%

---

5

[koara.lib.keio.ac.jp](http://koara.lib.keio.ac.jp)

Internet Source

<1%

---

6

[media.proquest.com](https://media.proquest.com)

Internet Source

<1%

---

7

[tci-thaijo.org](http://tci-thaijo.org)

Internet Source

<1%

---

8

[alexleavitt.com](https://alexleavitt.com)

Internet Source

<1%

---

---

9

"ICT Critical Infrastructures and Society",

Springer Science and Business Media LLC,  
2012

Publication

<1%

---

10

Kafri, O.. "", Optical Engineering, 1986.

Publication

<1%

---

11

[pnrsolution.org](http://pnrsolution.org)

Internet Source

<1%

---

Yu Jiang. "Topic Sentiment Change Analysis", **12**  
Lecture Notes in Computer Science, 2011

Publication

<1%

---

13

[www.ukessays.com](http://www.ukessays.com)

Internet Source

<1%

---

14

[onlinelibrary.wiley.com](http://onlinelibrary.wiley.com)

Internet Source

<1%

---