**CHAPTER-1: INTRODUCTION**

Twitter is a “micro-blogging” social networking website that has a large and rapidly growing user base. Those who use Twitter can write short 140 characters long or less updates called ‘tweets’. ‘Tweets’ are seen by those who ‘follow’ the person who ‘tweeted’. Due to the growing popularity of the website, Twitter can provide a rich bank of data in the form of harvested “tweets”. Twitter by its very nature, allows people to convey their opinions and thoughts openly about whatever topic, discussion point or product that they are interested in sharing their opinions about. Therefore Twitter is a good medium to search for potentially interesting trends regarding prominent topics in the news or popular culture.

R studio is one of the many programmes that offer packages that can analysis data, however R studio works well with statistical problems and has a user friendly interface.

Sentiment analysis (or opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source material.

The value of Twitter in recent years has increased as businesses, political groups and curious Internet users alike have started to assess the public’s general sentiment for their products and services from twitter posts. Sentiment analysis provides a means of tracking opinions and attitudes on the web and determines if they are positively or negatively received by the public. The purpose of Text mining is to process unstructured (textual) information and to extract meaningful numeric indices from the text, allowing the application of various data mining algorithms to explain the textual dataset.

The classification model which this project will develop will determine whether the tweet status updates (which cannot exceed 140 characters) reflects positive opinion or negative opinion on the behalf of the person who tweeted. This paper will use a hybrid of knowledge based sentiment analysis methodologies which have been more traditionally used, and those of machine learning methodologies which used a more intuitive approach to sentiment. The results of these two methodologies will be used to perform a thorough analysis of the dataset.

**1.1 DOMAIN DESCRIPTION**

Twitter is different to other forms of raw data which are used for sentiment analysis, as sentiments are conveyed in one or two sentence blurbs rather than paragraphs. Twitter is much more informal and less consistent in terms of language. Users cover a wide array of topics which interest them and use many symbols such as emoticons to express their views on many aspects of their life. When using human generated status updates, sentiment are not always obvious. Many tweets are ambiguous and can use humour to maximize the opinion to other human readers but deflect the opinion to a machine learning algorithm. This provides a challenge for machine learning algorithms. Sentiment analysis provides a means of tracking opinions and attitudes on the web and determines if they are positively or negatively received by the public.

Sentiment analysis is usually conducted between two levels:-

**Coarse level: -** Coarse level sentiment analysis deals with determining the sentiment of an entire document and

**Fine Level: -** Fine level deals with attribute level sentiment analysis. Sentiment analysis in Twitter provides a dramatically different data set where multiple interesting challenges can arise.

**1.2 MOTIVATION**

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover the response on twitter is more prompt and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis). Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analysing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm’s stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favourable response and in which a negative response (since twitter allows us to download stream of geo-tagged tweets for particular locations. If firms can get this information they can analyze the reasons behind geographically differentiated response, and so they can market their product in a more optimized manner by looking for appropriate solutions like creating suitable market segments. Predicting the results of popular political elections and polls is also an emerging application to sentiment analysis.

**1.3 GOALS**

In order to conduct any kind of analysis on twitter the construction of a suitable dataset of tweets needs to be built. Twitter API is an app which extracts tweets from twitter and loads them into a dataset. The aims of this Twitter API are threefold:-

* To construct a database of tweets on the keywords #Farhan Akhtar which will be built using a Twitter API app.
* R studio will perform a series of analysis on the data such as a knowledge based techniques which uses a sentiment lexicon dictionary to determine the number of positive and negative tweets. Machine learning techniques which are based on a training set and will determine the number of tweets which are positive and negative.
* Use the results from the knowledge based techniques and those of the machine learning techniques to ensure a thorough analysis of the dataset.

**CHAPTER-2: SOFTWARE REQUIREMENTS**

The tweets about # Farhan Akhtar were collected from Twitter through twitter API. There were about 120 tweets about # Farhan Akhtar .The tweets were kept in a separate .csv file for further analysis. Some of the other requirements are: -

User Interfaces: - The programme used for this project was R studio. R studio was selected as a suitable programme as its interface and programme libraries met the requirements of the brief. Additionally the user interface of R studio was simple to navigate and easy to understand. R studio is free and open source, and works well on both Windows and Mac hardware. It contains advanced statistical routines as well as a large, coherent, integrated collection of tools for data analysis. R studio also processes powerful graphics capabilities which aid visualisation of data and results greatly. Due to these properties Control of the programme allowed the user to interact with the application at optimal ease.

The interface will include user inputs as well as two graphics, as outlined below**:**

1. User Inputs (Mandatory)

The user will be able to control the sentiment analysis of topics in two ways:

* **Edit** this function will let the user edit Keywords, by adding, editing, orremoving keywords for each topic, and
* **Time**, this function will let the user specify the duration of each analysissession.

1. Graphic 1: Topic Mood Gauge (Mandatory)

This graphic will consist of a simple histogram, which shows the current mood of the Twitter community on the topic of #FarhanAkhtar. The percentage of the Twitter users will be displayed who are currently for or against the topic being analyzed. It will also display the most frequently used words used on the subject of #FarhanAkhtar through the use of a word cloud.

1. Error Notifications (Mandatory)

Error notifications will be required within the programme R studio:- this will be presented to the user with appropriate messages in blue, which will describe the error that has taken place. If applicable, error messages suggest possible solutions to the problem.

1. Hardware Interface

The application will run on a password protected personal Microsoft laptop. No further hardware devices or interfaces will be required for this analysis.

1. Software Interfaces

* Inputs

The software will receive input from four sources.

1. The programme R studio: - The programme R studio will supply the code results and the majority of the graphs for the analysis.
2. The Twitter API: - The Twitter API will supply dataset of the Tweet text.
3. MS Excel: - It will hold the dataset once retrieved from Twitter API.
4. Twitter app’ Sentiment140’: - Sentiment140 app will supply an additional pie chart which will add a more visual element to the interpretation of the data.

* Outputs

The output will portray the current mood of the Twitter community on # FarhanAkhtar in the form of a word cloud.

It will show a pie chart depicting the opinion of people about the topic.

**2.1** **FUNCTIONAL REQUIREMENTS**

1. Retrieving Input

The software will receive three inputs: R studio code and R studio libraries, analysis session duration and Tweets.

* R studio code, R studio libraries was entered by the user for each topic.
* The analysis session duration will be set by the user before each session.
* Tweets will be retrieved from the Twitter API and saved on an Excel file.

1. Real-Time Processing

The software will take input, process data, and display output in real-time. This will ensure the data provided by Twitter is a current view of the Twitter community’s mood on # FarhanAkhtar.

1. Sentiment Analysis

Sentiment analysis will be performed on the keywords within the Tweet to determine the overall mood of the Tweet relative to the topic. The sentiment analysis will provide a negative or positive numeric sentiment value.

1. Output

The software must output real time data in the form of pie charts and word cloud. In addition, the software may output additional statistics pertaining to a topic of #FarhanAkhtar . This output will be clear and easy to understand.

* 1. **NON FUNCTIONING REQUIREMENTS**

1. Performance

* The Twitter API will provide up-to-date information:- limited only by the rate of Twitter input. R studio will provide promptly analysis of the data using the various software packages available to it. The output should display the latest results at all times, and if it lags behind, the user should be notified. The application should be capable of operating in the background should the user wish to utilize other applications.

1. Reliability

* The software will meet all of the functional requirements without any unexpected behaviour. At no time should the output display incorrect or outdated information without alerting the user to potential errors. In this instance error message will be shown.

1. Availability

* The software will be available at all times on the user’s device desktop or laptop, as long as the device is in proper working order. The functionality of the software will depend on any external services such as internet access that are required. If those services are unavailable, the user should be alerted.

1. Security

* The software should never disclose any personal information of Twitter users, and should collect no personal information from its own users. The use of passwords and API keys will ensure private use of the Twitter API. The programmes will be performed on a password protected laptop and desktop to ensure maximum security.

1. Maintainability

* The software should be written clearly and concisely. The code will be well documented. Particular care will be taken to design the software modularly to ensure that maintenance is easy.

1. Portability

* This software will be designed to run on any Windows or Linux or Mac operating system.

**2.3 LOGICAL DATABASE REQUIREMENTS**

* The tweets taken from Twitter will be stored on an excel spreadsheet. Excel is an excellent programme for storing large amounts of data as well as being easy to upload the data to R Studio .The data will have two columns, column one will have the score of the tweets ( positive, negative,) column two will store the actual tweet content. Each row will represent an individual tweet.

**CHAPTER-3:PROJECT DESIGN**

The construction of a suitable dataset of tweets needs to be built. Twitter API is an app which extracts tweets from twitter and loads them into a dataset. Additionally for this project R studio was used along with the following packages and libraries:

* TwitteR **:-** Provides an interface to the Twitter API
* ROAuth **:-** This package provides an interface to the OAuth 1.0 specification,which allows users to authenticate via OAuth to the server of their choice
* Plyr :- The package Plyr is a set of tools that solves common problems bybreaking down bigger problems into more workable pieces. The package then operates on each problem before reassembling the reworked pieces back together.
* Stringr **:-** Stringr makes R string functions more consistent, simpler and easier touse by ensuring the function and argument names are consistent and all functions deal with NA’s and zero length characters appropriately. Stringr also ensures that the data output from each function matches the input data structures of other functions.
* Ggplot2 :- This package provides an implementation of the grammar of graphicsin R, combining the advantages of both base and lattice graphics. Plots can be built up step by step from multiple data sources
* RColourBrewer :- This packages provides palettes for drawing maps shadedaccordingly to a variable

# **3.1 DESIGN STRUCTURE**

* Related work describes the related literature that has been reviewed around the topic of sentiment analysis and the various methodologies used. The related work is loosely divided up into knowledge based approaches and machine learning based approaches.
* Design and Architecture describes how the overall layout of the project will be achieved. The design of the paper is based on Naive Bayes classifier which is explained in this section.
* The Architecture diagram provides a ‘road map’ of the different distribution systems that will be used and how they are connected to other systems.
* Implementation provides a step by step guide through the processes of the paper which includes the code that was used as well as outputs.
* Requirements and specifications provide an insight into the aspects of the project that must be considered such as Functional requirements, Data requirements, User requirements etc.

# **3.2 DATASETS**

A Twitter API app was used to pull tweets from Twitter's public timeline in real-time. A dataset was created using twitter tweets from a topic that was dominating twitter at the time of data collection:- #Farhan Akhtar .But our concentration were mainly to the tweets returned from # Farhan Akhtar which returned 120 tweets since the other topic did not fetch much tweets. A sentence level sentiment analysis was performed on tweets as many were full of slang words and misspellings. This is done in three phases. In the first phase of a sentence level sentiment analysis pre-processing is done. Secondly a feature vector is created using relevant features. A publicly available sentiment lexicon which consists of around 6800 words in a [list of positive and](http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar) [negative opinion words or sentiment words for English was used to separate the](http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar) [tweets.](http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar)  The final sentiment is based on the number of tweets in each class using several sentiment analysis methodologies:- the bag-of-words approach, which uses available lexical resources sentiment analysis. Machine learning approaches are also used where the tweets dataset was split in two Training and testing. These tweets were then used for training and testing so to conduct a Naive Bayes classifier.

**3.2.1 CREATION OF A DATASET**

Since standard twitter dataset are not available for analysis, we created a new dataset by collecting tweets over a period of time. Tweets were collected automatically using Twitter API and they are manually annotated as positive or negative. In total 120 tweets were collected from # Farhan Akhtar .Unexpectedly a number of the tweets were neutral, however positive and negative tweets created the dataset.

* PREPROCESSING OF TWEETS

1. Pre-processing steps were performed in order to ensure that Keyword extraction was made as simple as possible for the algorithm.
2. Punctuation marks, correctors and digits were removed as well as changing the tweet texts to lower case, splitting sentences to words with structural split and comparing the corpus from the dictionary’s positive and negative words. The matched term would be returned as a true or false value which will be treated as 1 or 0 by the sum function. Finally scores are put into a data frame named scores.df . Before tweets can be scores however a sentiment lexicon of words must be obtained. The final score for each tweet will be the number of positive words minus the number of negative words. If the score is higher than 0, the tweet will be regarded as positive. If the tweet score is lower than 0 the tweet will be regarded as negative opinion.

# **UML DIAGRAMS**

**3.3.1 USE-CASE DIAGRAM**

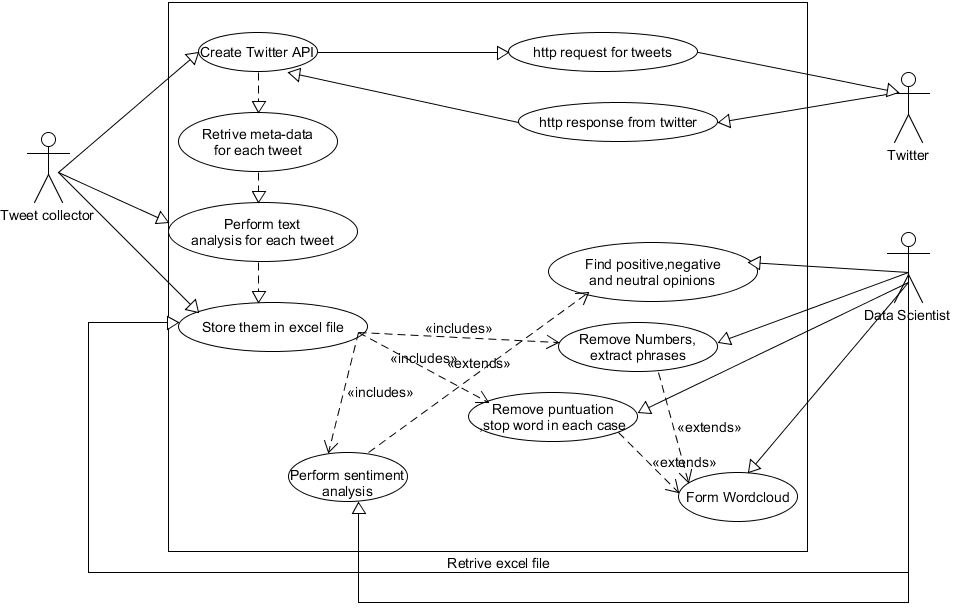


Figure 3.3.1: - Use Case Diagram.

**Description: -**

* As it is evident from the use case diagram, here the twitter collector which in our case it the Twitter API is going to collect the tweets from the twiiter database, but first the person who is going to collect the tweets and store them in the database have to login to his twiiter account and create one application.
* A sort of handshaking process goes between the twiiter API and the twitter collector who is collecting the tweets in his PC.The collector is sending http request to the twitter through his twitter API, and twiiter is sending http response message along with the extracted tweets about the topic asked.
* Data scientist on the other hand, remove unwanted English stop words, punctuations, numbers etc.
* Further, he performs sentiment analysis to it and finally derives the percentage of positive, negative and neutral opinion of people.
  + 1. **ACTIVITY DIAGRAM**

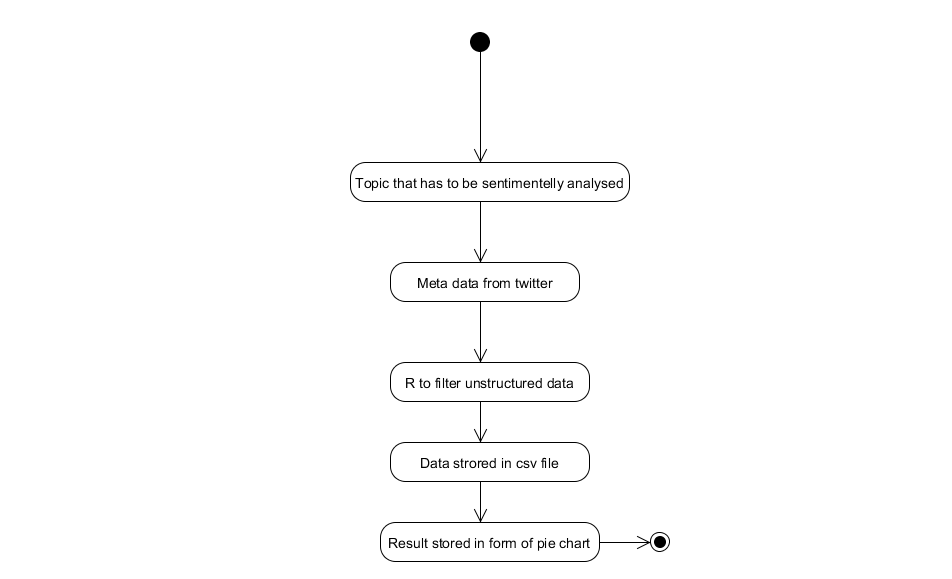


Figure 3.3.2: - Activity Diagram.

**Description: -**

* As the above figure is showing, first step is to choose the topic of sentiment analysis.
* 2nd step extracts the Meta data from the twitter through the Twitter API about that topic.
* The data extracted will be totally unstructured and will be in the form of list, In that case we used R to convert the unstructured data into structured form.
* Finally the data which is now totally structured and ready to be analysed is stored in an Excel file.
* At last, all the required operations necessary were executed by the data scientist and finally the results were depicted in the form of pie chart.

**3.3.3 SEQUENCE DIAGRAM**

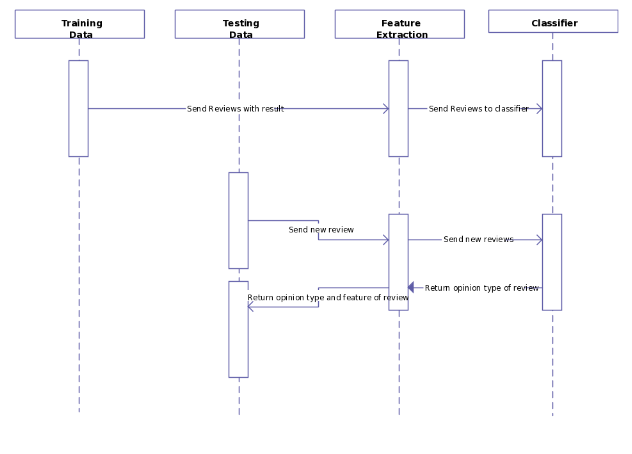


Figure 3.3.2: - Sequence Diagram.

**Description: -**

* Training data is the data on which the machine learning programs learn to perform correlational tasks (classify, cluster, learn the attributes)
* Testing data is the data, whose outcome is already known (even the outcome of training data is known) and is used to determine the accuracy of the machine learning algorithm, based on the training data (how effectively the learning happened).
* There is no rule of thumb for selecting training / testing data set sizes, but there is a broad agreement among practitioners that 10% of the total data should be training data, the next 25% should be testing data. Generally speaking, testing data set should at least be twice than that of training data set.

**CHAPTER-4: INTRODUCTION TO R PROGRAMMING LANGUAGE**

**4.1 R ENVIRONMENT OVERVIEW**

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

* + an effective data handling and storage facility
  + a suite of operators for calculations on arrays, in particular matrices
  + a large, coherent, integrated collection of intermediate tools for data analysis
  + graphical facilities for data analysis and display either directly at the computer or on hardcopy,

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages.

* 1. **RELATED SOFTWARE AND DOCUMENTATION**

R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. The evolution of the S language is characterized by four books by John Chambers and co-authors. For R, the basic reference is The New S Language: A Programming Environment for Data Analysis and Graphics by Richard A. Becker, John M. Chambers and Allan R. Wilks. The new features of the 1991 release of S are covered in Statistical Models in S edited by John M. Chambers and Trevor J. Hastie. The formal methods and classes of the methods package are based on those described in Programming with Data by John M. Chambers. There are now a number of books which describe how to use R for data analysis and statistics, and documentation for S/S-Plus can typically be used with R, keeping the differences between the S implementations in mind. See Section “What documentation exists for R?” in The R statistical system FAQ.

**4.3 R AND STATISTICS**

Our introduction to the R environment did not mention statistics, yet many people use R as a statistics system. We prefer to think of it of an environment within which many classical and modern statistical techniques have been implemented. A few of these are built into the base R environment, but many are supplied as packages.

There are about 25 packages supplied with R (called “standard” and “recommended” packages) and many more are available through the CRAN family of Internet sites (via https://CRAN.R-project.org) and elsewhere. Most classical statistics and much of the latest methodology is available for use with R, but users may need to be prepared to do a little work to find it.

In S a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give copious output from a regression or discriminant analysis, R will give minimal output and store the results in a fit object for subsequent interrogation by further R functions.

* 1. **R AND THE WINDOW SYSTEM**

The most convenient way to use R is at a graphics workstation running a windowing system. This guide is aimed at users who have this facility. In particular we will occasionally refer to the use of R on an X window system although the vast bulk of what is said applies generally to any implementation of the R environment. Most users will find it necessary to interact directly with the operating system on their computer from time to time. In this guide, we mainly discuss interaction with the operating system on UNIX machines. If you are running R under Windows or mac OS you will need to make some small adjustments. Setting up a workstation to take full advantage of the customizable features of R is a straightforward if somewhat tedious procedure, and will not be considered further here. Users in difficulty should seek local expert help.

* 1. **USING R INTERACTIVELY**

When you use the R program it issues a prompt when it expects input commands. The default prompt is ‘>’, which on UNIX might be the same as the shell prompt, and so it may appear that nothing is happening. However, as we shall see, it is easy to change to a different R prompt if you wish. We will assume that the UNIX shell prompt is ‘$’. In using R under UNIX the suggested procedure for the first occasion is as follows:

1. Create a separate sub-directory, say work, to hold data files on which you will use R for this problem. This will be the working directory whenever you use R for this particular problem. $ mkdir work $ cd work

2. Start the R program with the command $ R

3. To quit the R program the command is > q() At this point you will be asked whether you want to save the data from your R session. On some systems this will bring up a dialog box, and on others you will receive a text prompt to which you can respond yes, no or cancel to save the data before quitting, quit without saving, or return to the R session.

Data which is saved will be available in future R sessions. Further R sessions are simple.

1. Make work the working directory and start the program as before: $ cd work $ R

2. Use the R program, terminating with the q() command at the end of the session.

To use R under Windows the procedure to follow is basically the same. Create a folder as the working directory, and set that in the Start In field in your R shortcut. Then launch R by double clicking on the icon.

# **CHAPTER-5: IMPLEMENTATION**

## **5.1 Creating a Twitter Application**

First step to perform Twitter Analysis is to create a twitter application. This application allows the performance of sentiment analysis by connecting your R console to the twitter using the Twitter API. The steps for creating the twitter applications are as follows:

* Go to https://dev.twitter.com and login by using a twitter account.
* Then go to My Applications and create a new application.

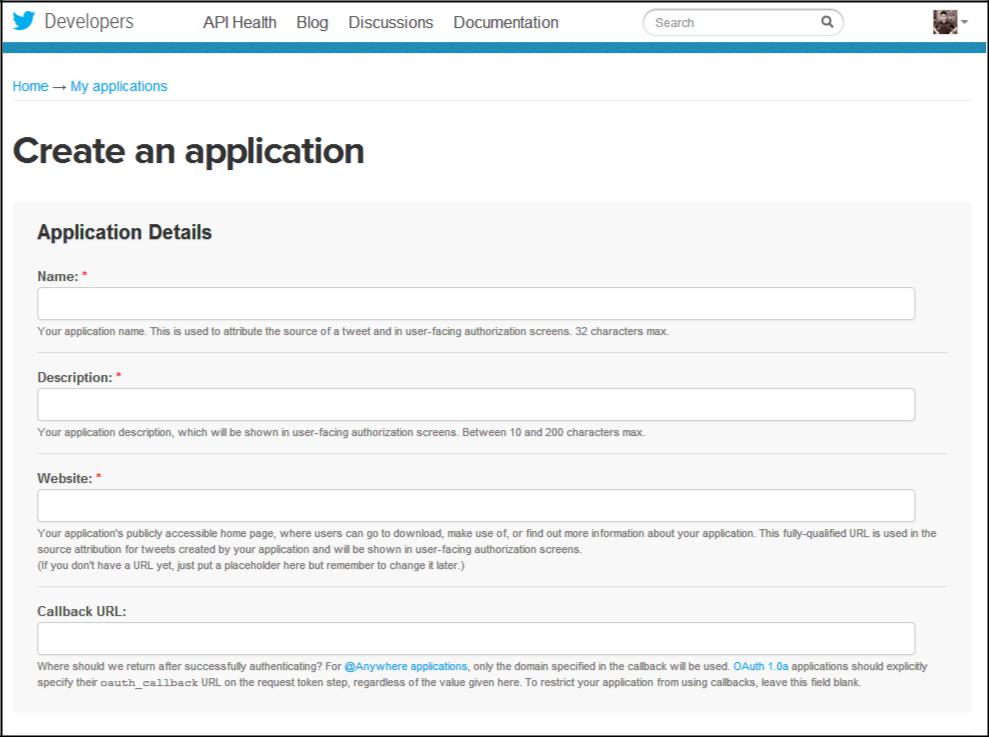


Figure 5.1.1: - Layout of Twitter Layout

Give an application name (in this case komrush1), describe about your application (in this case twitter API), provide a website’s URL (student email URL). The Call-back URL was left blank. Complete other formalities and create a twitter application. Once, all the steps are done, the created application will show as below. The properties were changed to ‘Read Write and Access Direct Messages’.

The Consumer key and Consumer Secret numbers were used in R Studio.

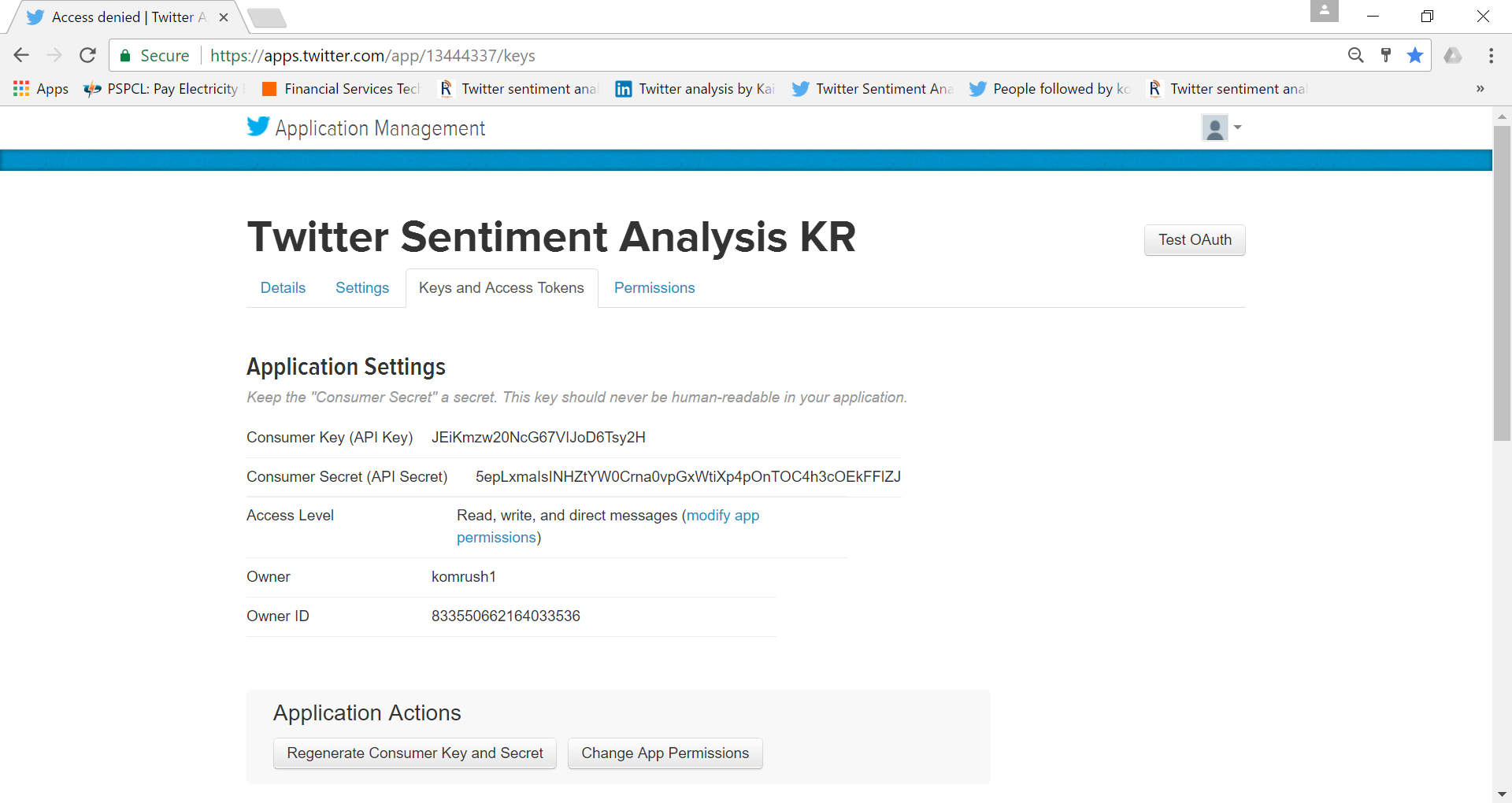


Figure 5.1.2: - Our twitter API

**5.2 WORKING ON R STUDIO**

* The installation of some packages and libraries in R was the first step of the R studio process. These are twitter, ROAuth, plyr, stringr and ggplot2. The installation of these packages can be seen below by the following commands:

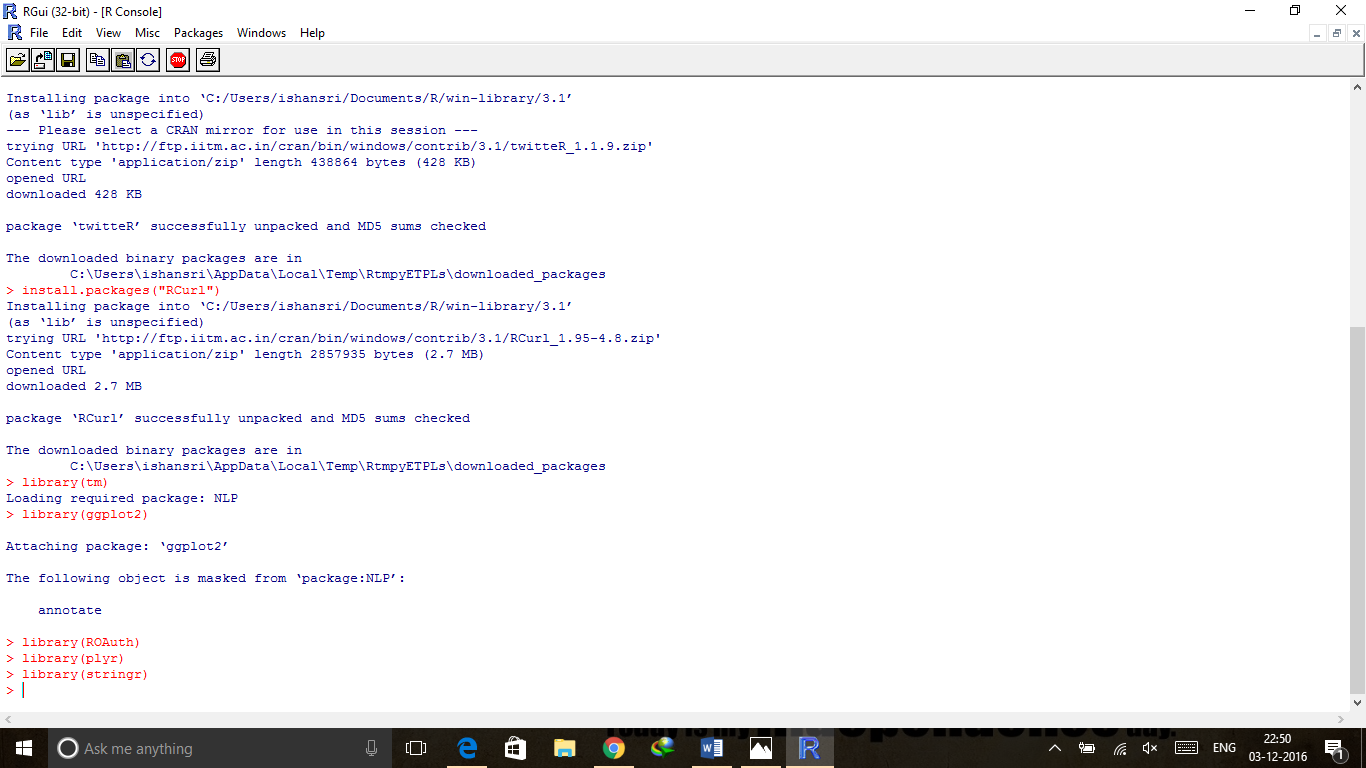


Figure 5.2.1: - The figure denotes the packages used.

* The next stage is to access the Twitter API. This step includes the script code to perform handshake using the Consumer Key and Consumer Secret number of the application. In figure you see the code you have to run to perform handshake.

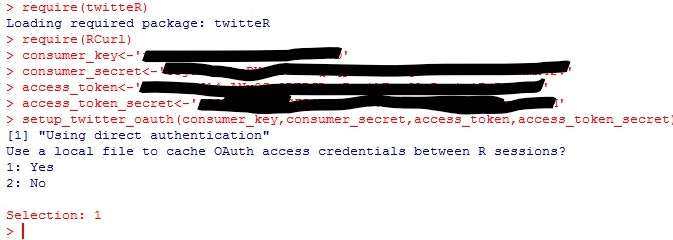


Figure 5.2.2: -The authorization process of API with twitter.

* In order to access the Twitter API, the programme requires the request URL, access URL and authorization URL of Twitter application to the variables request URL, access URL and authorization URL respectively. Consumer Key and consumer Secret are unique to a twitter application. It’s a sort of handshaking process by which twitter authorizes the user to access its data through this API.

## **5.3 SAVING TWEETS**

* Once the handshake is done and authorized by twitter, we can fetch most recent tweets related to any keyword. The code for getting tweets related to #FarhanAkhtar:

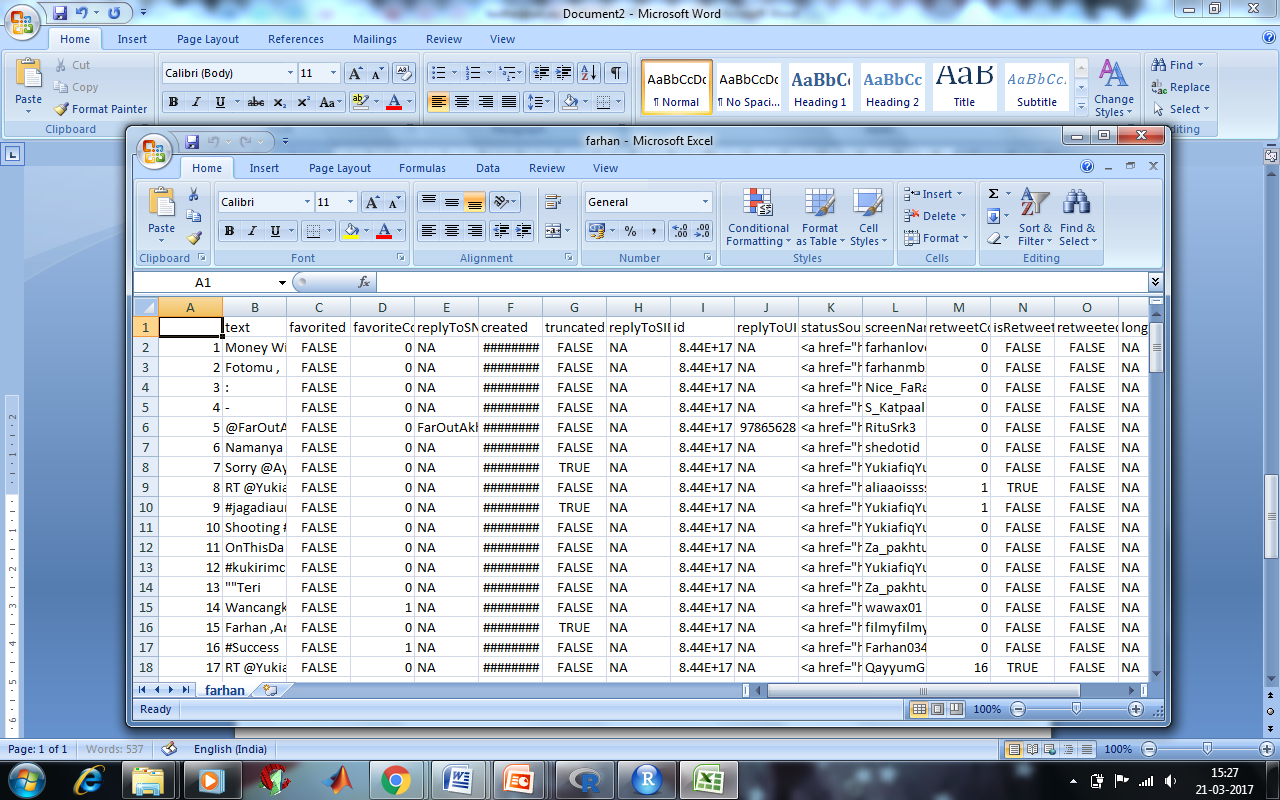


Figure 5.3.1: - The database that was created.

* This command will get 120 tweets related to ##FarhanAkhtar. The function “searchTwitter” is used to download tweets from the timeline. Twitter API can only return a fixed maximum amount of tweets (120). The return of a maximum number of tweets may not be met sometimes as there are not enough tweets for the particular keyword. Then we downloaded the Tweets for ##FarhanAkhtar, the results did not return many tweets, therefore it was decided that the paper would concentrate all of its efforts on the tweets pulled by ##FarhanAkhtar. As can be seen in the code above, the data of 120 tweets was converted into a data frame, so that analysis can be performed on it. Finally the data was converted into a .csv file.

## **5.4 SENTIMENT FUNCTION**

* Once the tweets were obtained, the application of some functions to convert these tweets into some useful information was needed. The main working principle of sentiment analysis is to find the words in the tweets that represent positive sentiments and find the words in the tweets that represent negative sentiments. For this a list of words that contains positive and negative sentiment words was needed. A list of positive and negative words complied which was publicly available was downloaded from Twitter. After downloading the list, it was saved it in a working directory. The sentiment analysis uses two packages **plyr** and **stringr** to manipulate strings.
* The sentiment function calculates score for each individual tweet. It first calculate the positive score by comparing words with the negative words list and then calculate negative score by comparing words with negative words list. The final score is calculated as

**Score = positive score – negative score.**

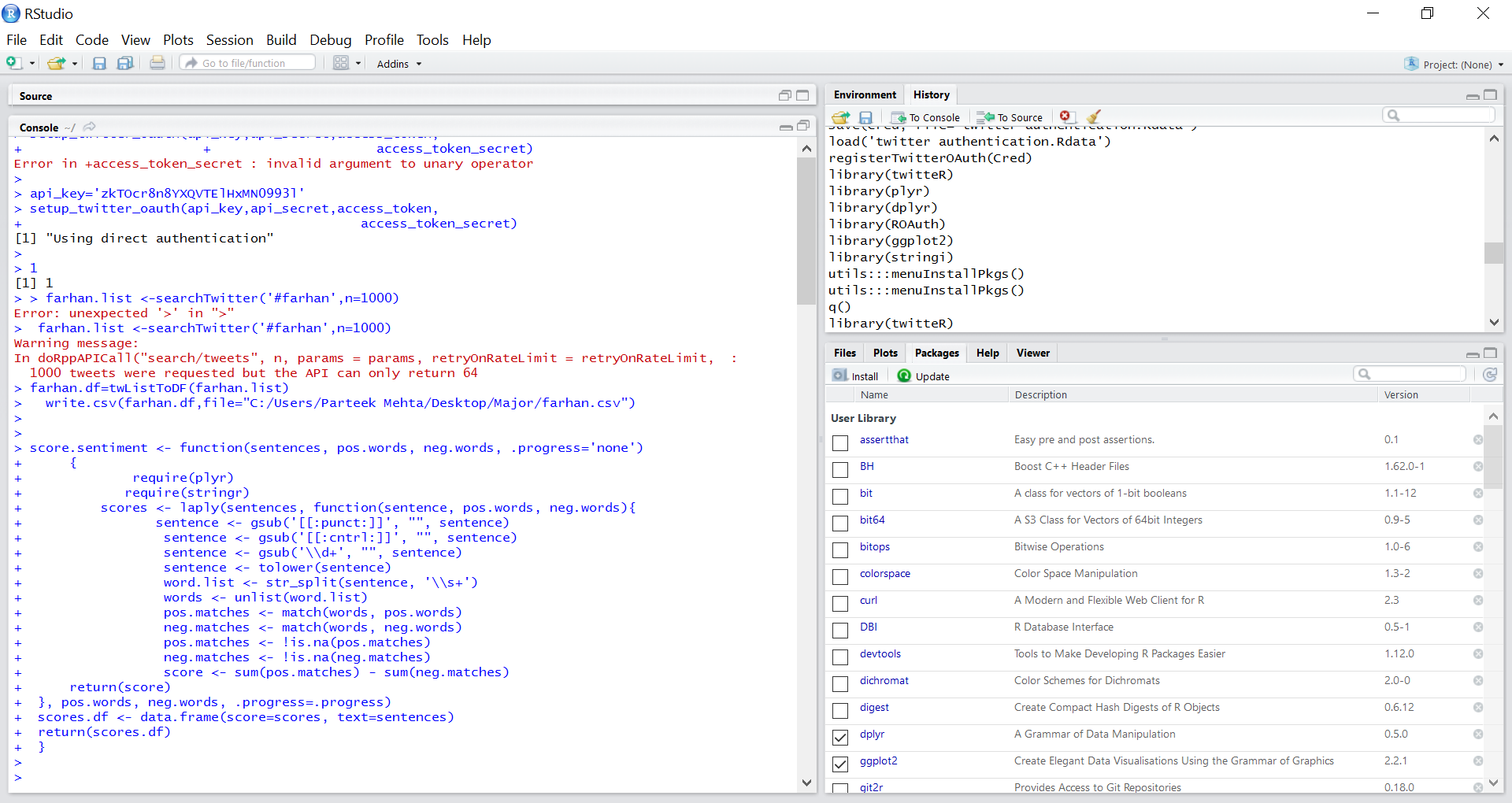


Figure 5.4.1: - Code for Sentiment analysis.

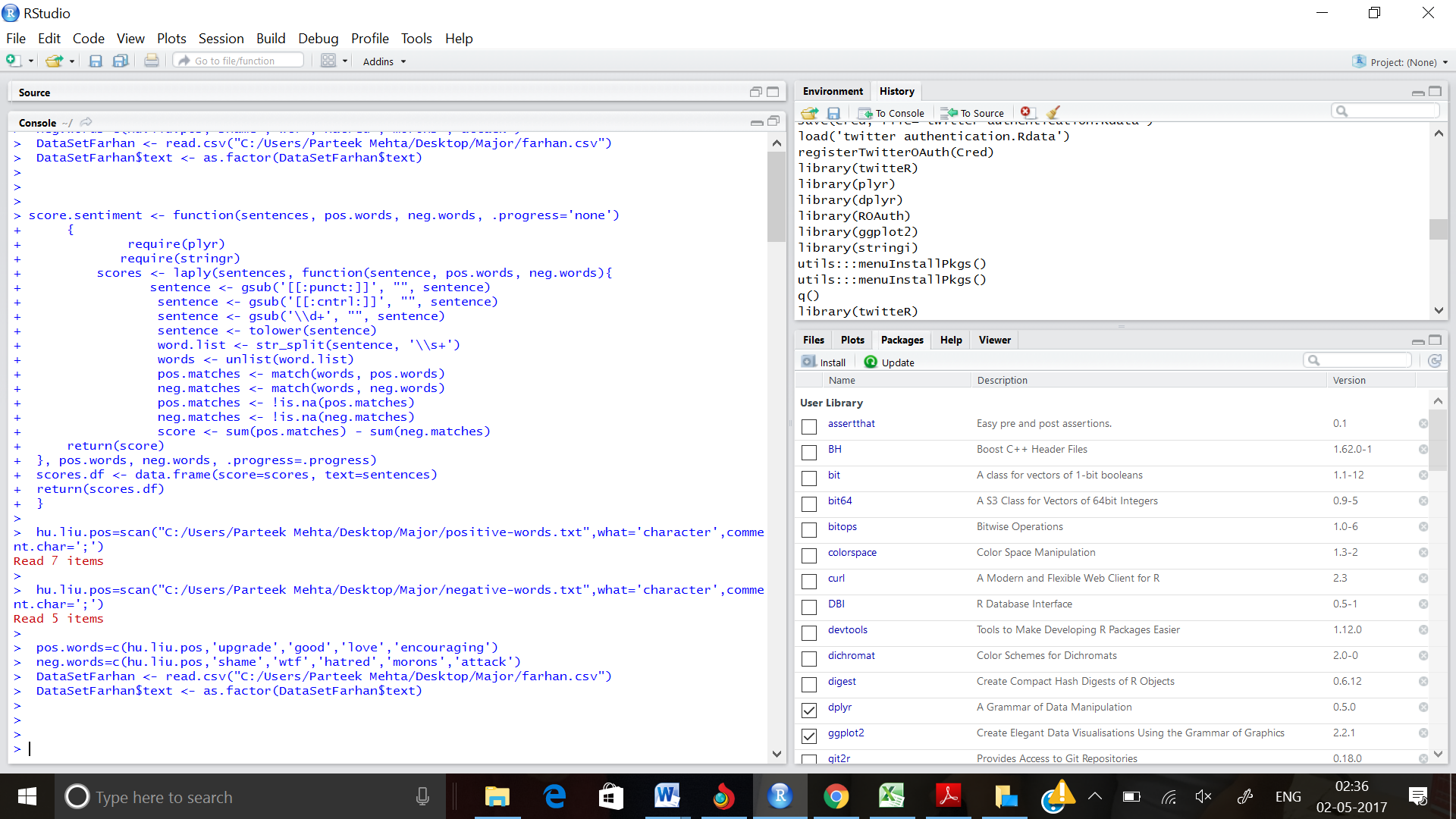


Figure 5.4.2: - The code for positive and negative tweets creation.

* The Figure above depict the negative columns that has been stored in our csv file. We need to find sum of all of that to get the final total negative value. The same steps can be done for positive as well as neutral tweets.

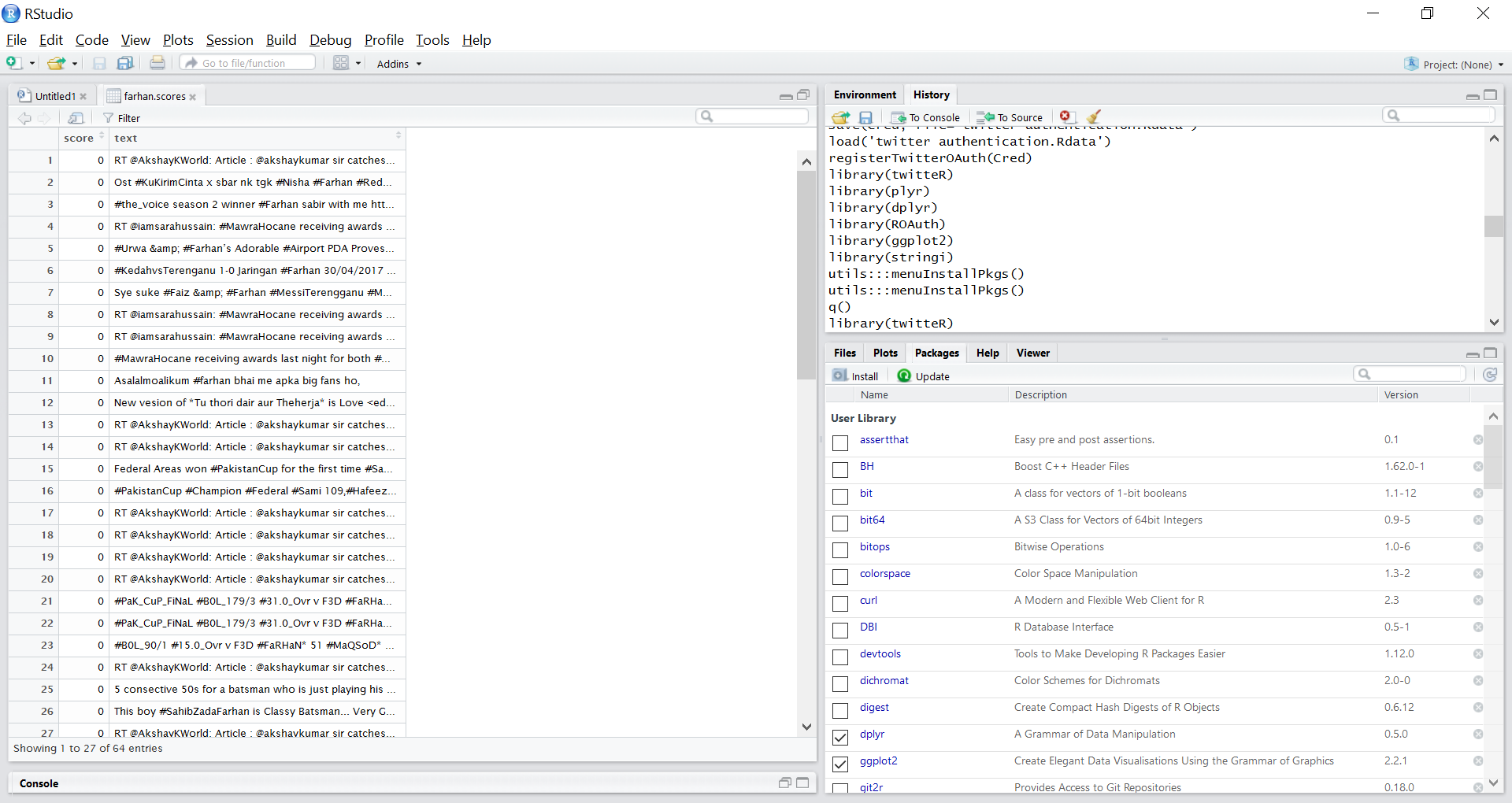


Figure 5.4.3: - The code for viewing the scores of farhan.

* These are some formulas that we have used to find total percentage of positive negative and neutral tweets aggregation is a vector which contains the sum of all elements of positive, negative and neutral columns.
* Pct variable contains the percentage of positive, negative and neutral opinions. As we can see no. of positive opinions are 50% negative 25% and neutral 25 %.

## 

Figure 5.4.4: - The histogram that was finally formed.

* Finally a histogram was constructed to present a better scenario of our analysis. As we can see the red area is depicting the people having positive opinion about the topic, Green depicts the negative opinion and blue depicts the positive opinion.

**CHAPTER-6: CONSTRAINTS**

* Twitter API has some limitations such as Twitter API can only return a fixed maximum amount of tweets (120). The return of a maximum number of tweets may not be met sometimes as there are not enough tweets for the particular keyword.
* Some tweets that we analysed were really few in numbers like around 100-150, so it was quite difficult to get the precise opinion of people about it.
* Twitter package was not available in the newer version of R, therefore we had to make our own package to meet our requirements. Twitter package is used to successfully authorize the user to connect itself to the twitter by giving his/her credentials.
* Same goes for the sentimental package, it was removed from the CRAN window of R so we had to make our own package to get the opinions.

**CHAPTER 7: FURTHER DEVELOPMENT OR RESEARCH**

The applicability of sentiment analysis for future businesses and marketing in using a keywords and analysis of the sentiments around that keyword by the public is only going to increase as the popularity of Twitter grows over the next few years. However, in terms of long-term development or research, the ability of the twitter API to pull data that is older, should be developed as well as other social media API’s so that sentiment analysis could be performed over a period of time, especially in the realm of social sciences where researchers could enquire into social and political shifts of opinion on the social media sites. Equally the lack of change in opinion over time on some issues might be worth pursuing as a topic of research for twitter sentiment analysis. The usefulness of such a sentiment analyser would allow for an interesting analysis of social and political issues.

**CHAPTER-8: CONCLUSION**

* In this paper a hybrid of knowledge based methodologies and machine learning methodologies were used in order to give a thorough examination of the tweets of #FarhanAkhtar which were extracted from Twitter.
* #FarhanAkhtar did not produce very many tweets therefore the concentration of the project fell to the tweets pulled from # FarhanAkhtar, which returned 120 tweets. Tweets were then classified into positive and negative classes using the machine learning classifier Naïve Bayes.
* The extraction of tweets from twitter proved to be more difficult than expected and several attempts were made to produce a dataset.
* Classification accuracy of the feature vector is tested using classifier like Nave Bayes. The assumption of Naïve Bayes that the data is independent, proved classification methodology to be an excellent tool in this analysis. It was found by the author that Machine learning algorithms were simpler to implement and more efficient than other aspects of the paper as they produced a table which allowed for transparency in the accuracy of the Naive Bayes classification. Overall the hybrid approach to sentiment analysis allowed for a thorough analysis of the data and performs well for a Twitter dataset. However, the accuracy of the Naïve Bayes classifier still leaves room for improvement this may be achieved by better pre-processing.

**REFERENCES**

**[1] Cambria, E; Schuller, B; Xia, Y; Havasi, C (2013).**["New avenues in opinion mining and sentiment analysis"](http://ieeexplore.ieee.org/document/6468032/). IEEE Intelligent Systems.

**[2] Roebuck, K. (2012-10-24)***.* [Sentiment Analysis: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors](https://books.google.com.br/books?id=kqsNBwAAQBAJ)

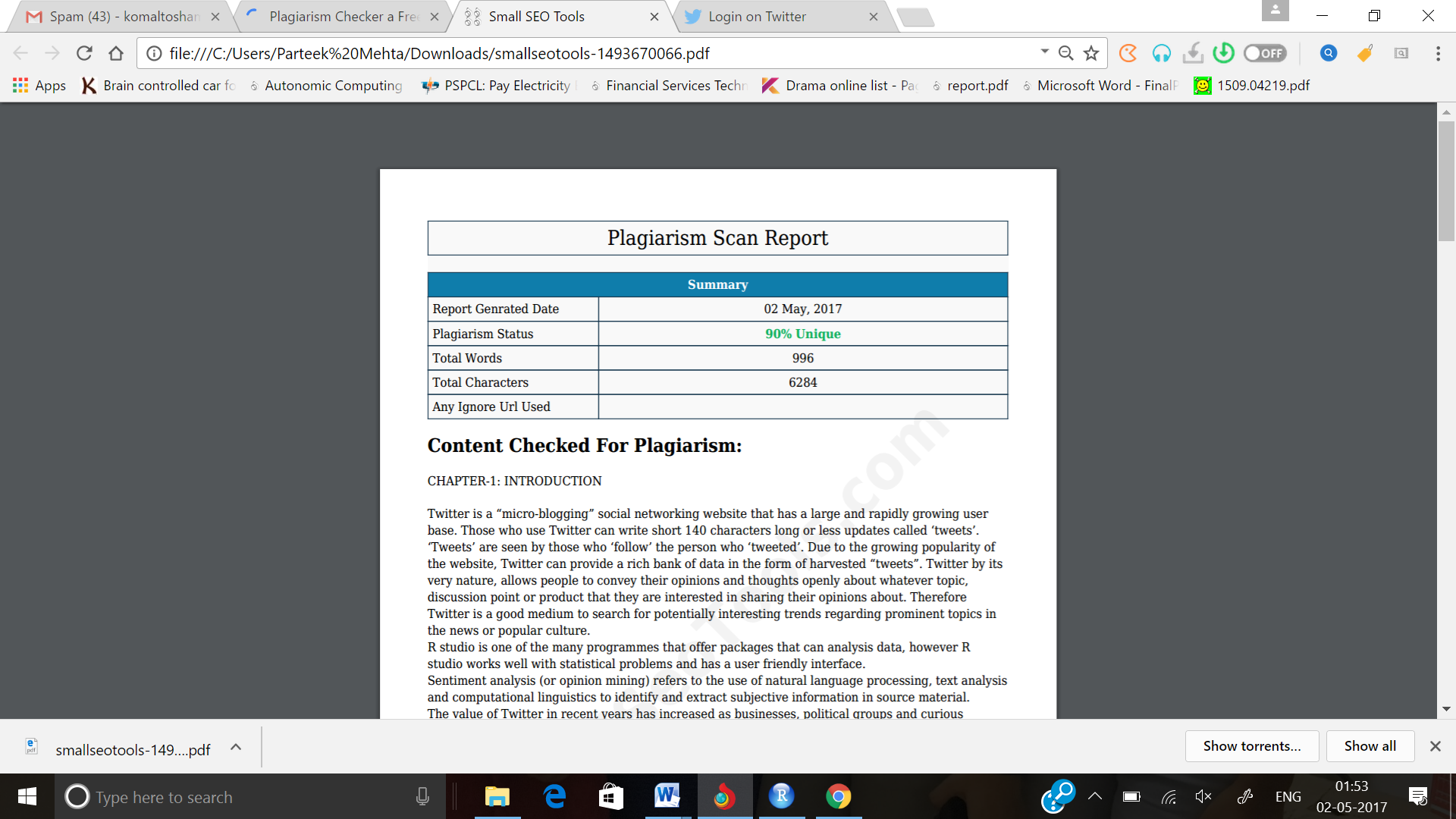
**[3]** <http://www.tutorialspoint.com/r/>

**[4]** <https://cran.r-project.org/bin/windows/base/>

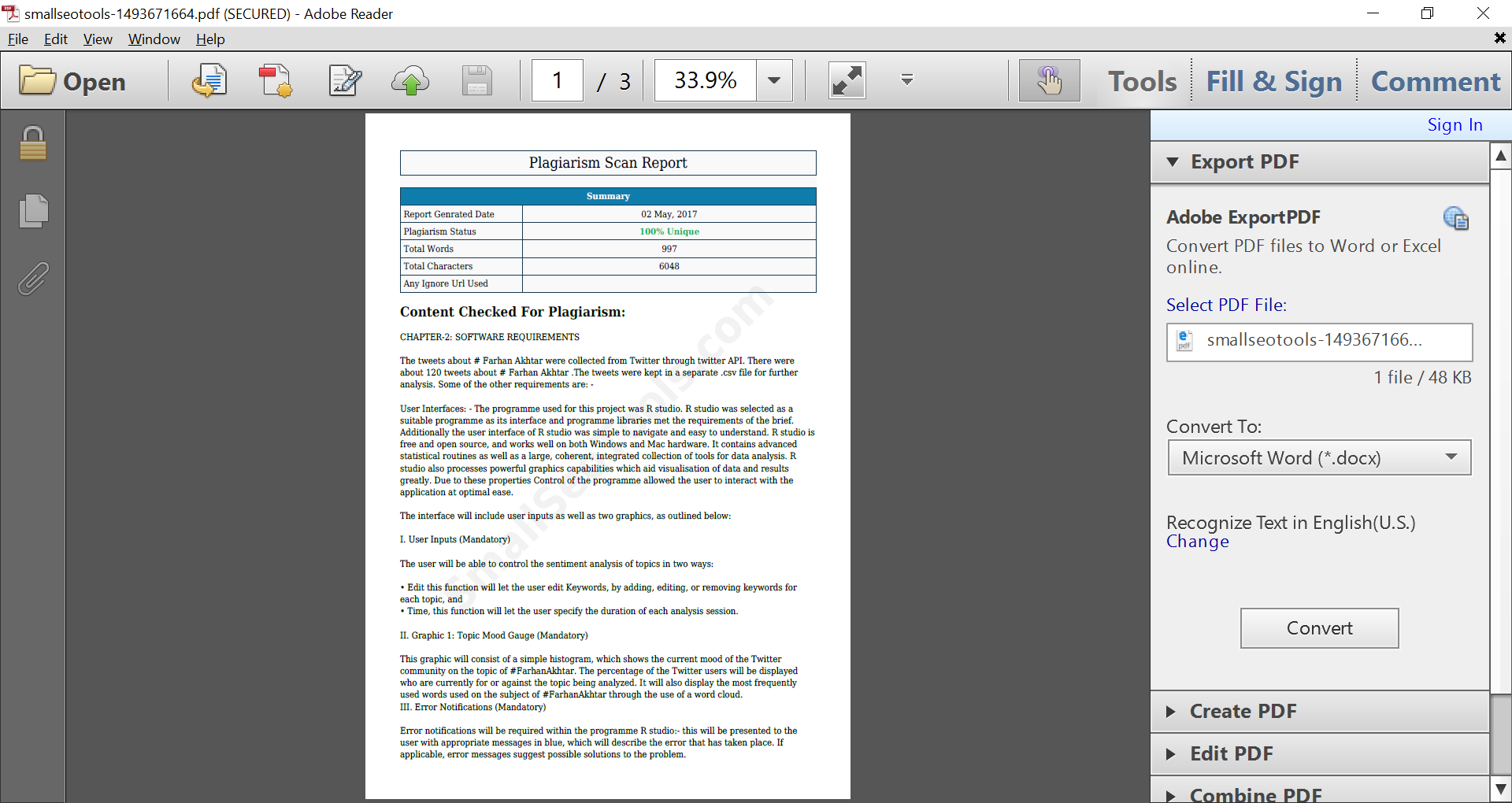
**[5]** <https://www.rstudio.com/products/rstudio/download/>

**PLAGIARISM REPORT**

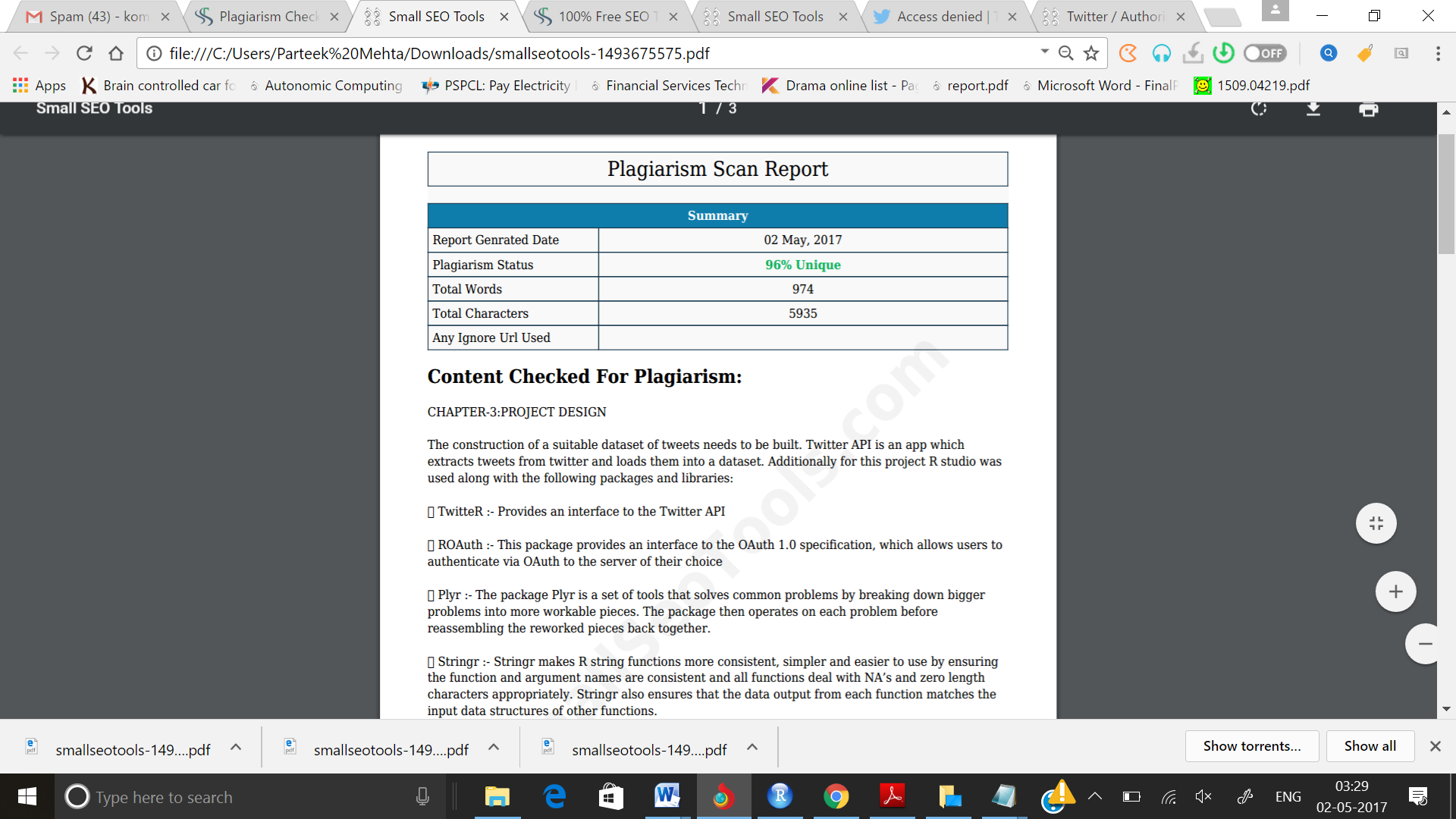
**CHAPTER-1 PLAGIARISM REPORT**



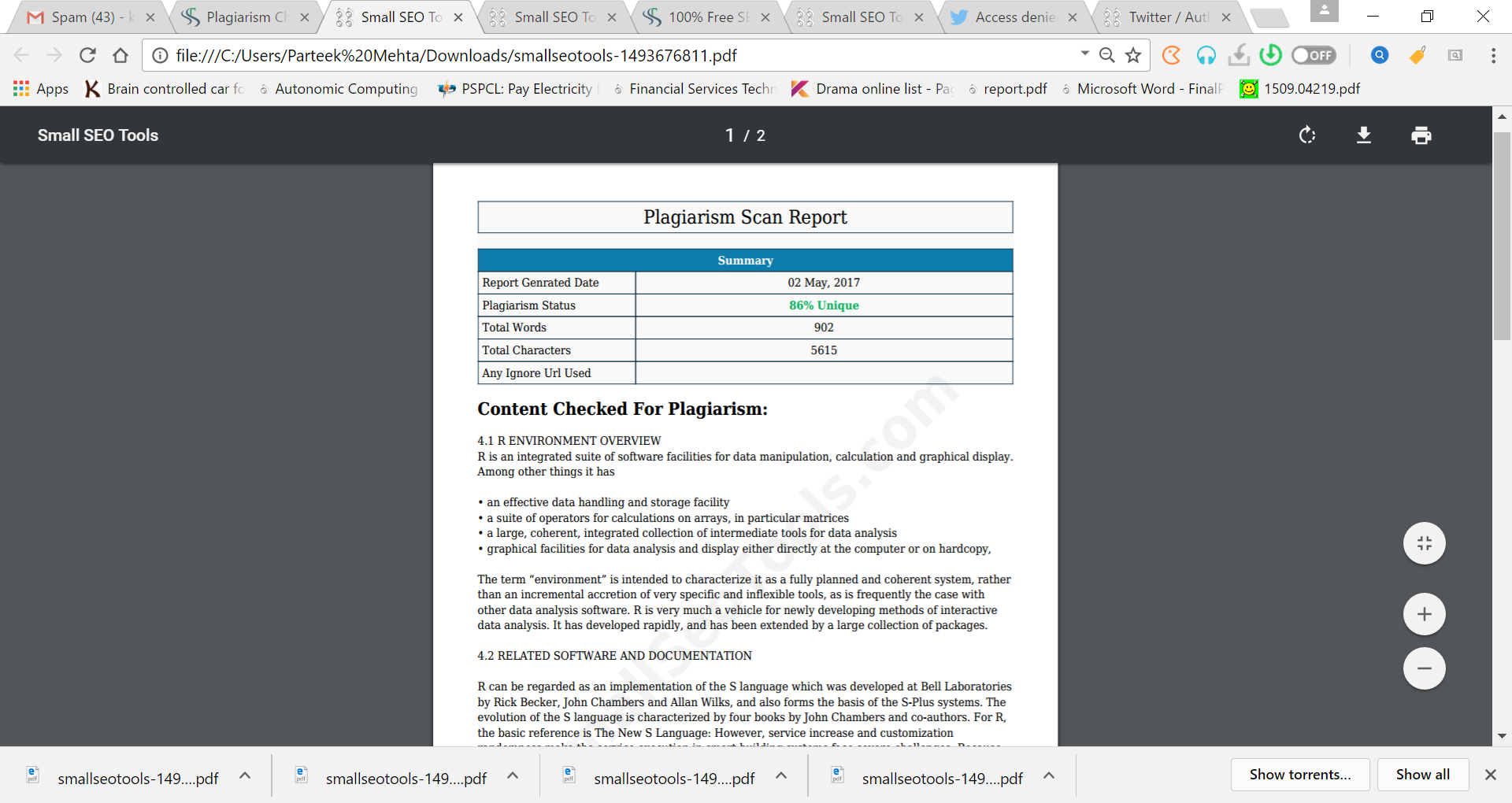
**CHAPTER-2 PLAGIARISM REPORT**



**CHAPTER-3 PLAGIARISM REPORT**



**CHAPTER-4 PLAGIARISM REPORT**



**GANTT CHART**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Month1 | Month2 | Month3 | Month4 | Month5 | Month6 |
| Project Briefing |  |  |  |  |  |  |
| Stage 1: Study of R Programming Language |  |  |  |  |  |  |
| Stage 2: Installation of R Studio and R |  |  |  |  |  |  |
| Stage 3: Registration and Authentication to Extract Tweets |  |  |  |  |  |  |
| Stage 4: Creation of csv file to Save Tweets |  |  |  |  |  |  |
| Stage 5: Check for Twitter Function and Apply it |  |  |  |  |  |  |
| Stage 6:Checking for any errors |  |  |  |  |  |  |
| Presentation and Report Submission |  |  |  |  |  |  |