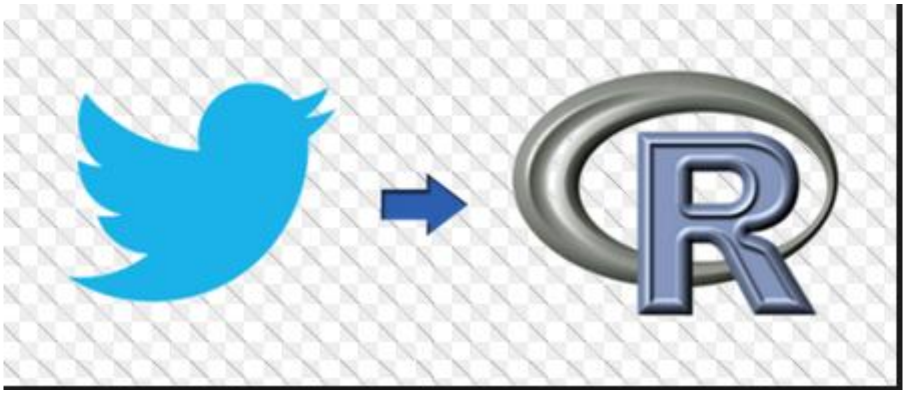
CIS 660

PROJECT FINAL REPORT



<u>**GROUP MEMBERS:**</u>

YASEEN AHMED SHAIK (2668256)

SAIYAM KOHLI (2669077)

PREFACE

This report documents the work done during product development lab in Cleveland State University under the supervision of Dr. Sun Sunnie Chung.

The report first shall give an overview of Twitter Mining using R tool and visualization and then about each part in detail. Report shall also elaborate on the future works which can be persuaded as an advancement of the current work.

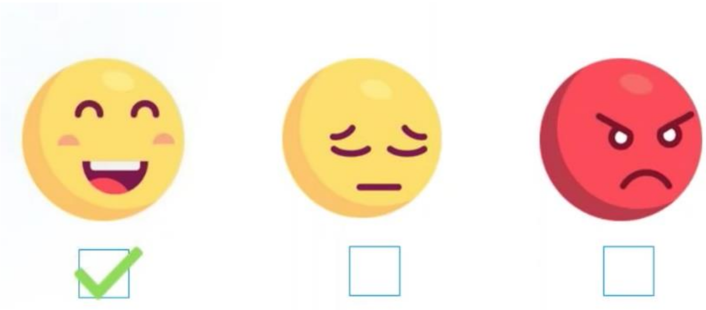
INTRODUCTION

● Numerous outlets available for individuals to express opinions and emotions...positive, negative, and neutral.
● Need to promote positive news, react to the negative, and move the needle favorably on neutral news....as near real-time as possible
● Mining high volume, high velocity data for meaningful insights is not easy!...too much, too fast
● Similar challenges exist across all industries/verticals

WHAT IS TWITTER SENTIMENT ANALYSIS?

Twitter is an online news and social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them.
Hence Twitter is a public platform with a mine of public opinion of people all over the world and of all age categories.

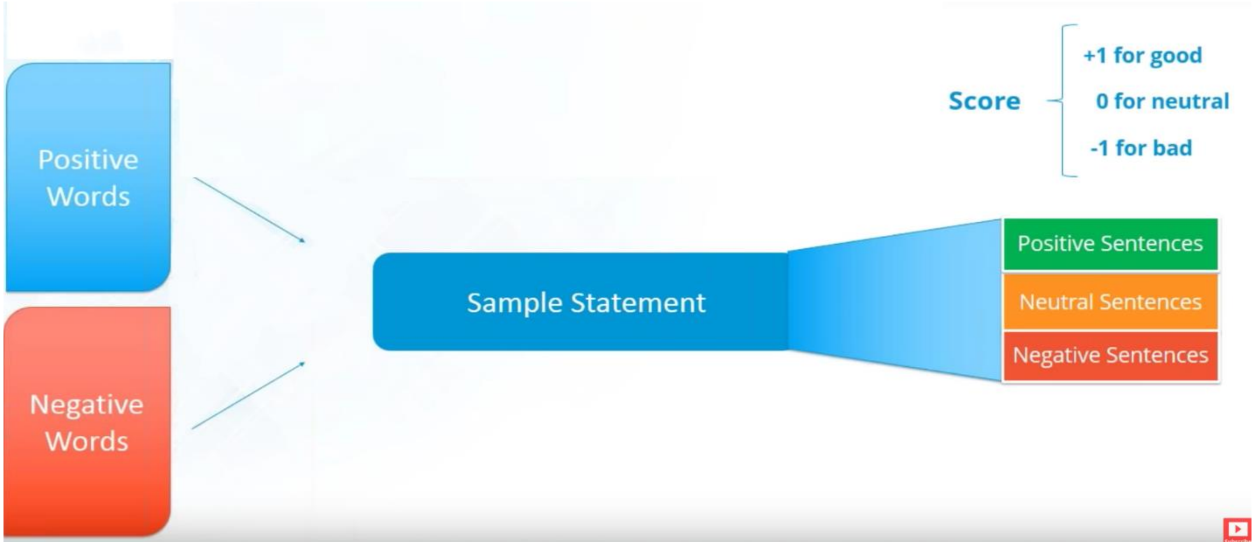
As of October 2016, Twitter has more than 315 million monthly active users .
Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the the attitudes, opinions and emotions expressed within an online mention.



WHY TWITTER SENTIMENT ANALYSIS?

The applications for sentiment analysis are endless. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics However, it is also practical for use in business analytics and situations in which text needs to be analyzed.
Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment in seconds, compared to the hours it would take a team of people to manually complete. Because it is so efficient (and accurate – Semantria has 80% accuracy for English content) many businesses are adopting text and sentiment analysis and incorporating it into their processes.

HOW DOES IT WORKS?



SYSTEM REQUIREMENTS

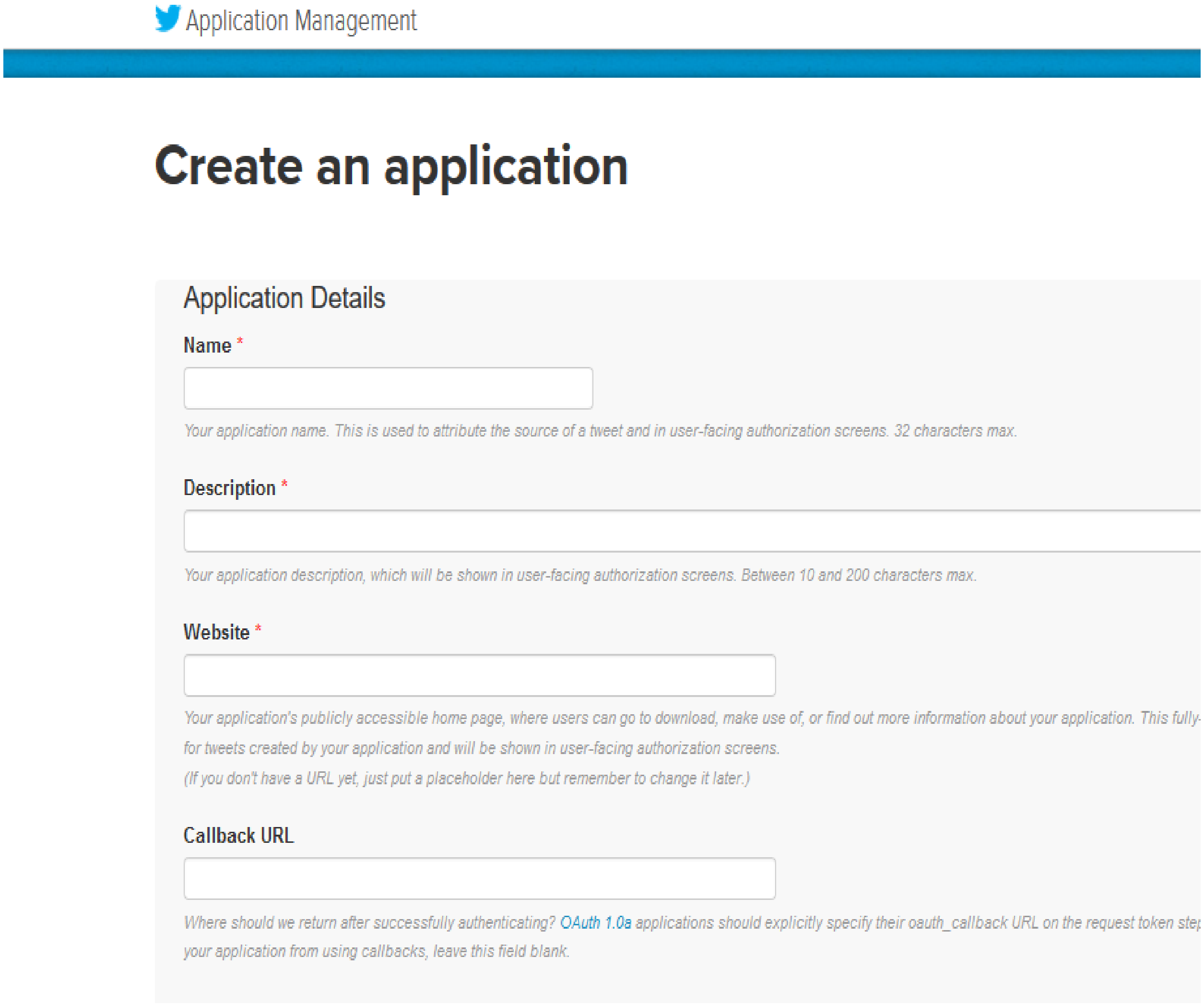
- Installation of R
- Twitter Authentication to access API

PROCESS

- **Initialization**: For our project twitter mining using R. We will choose a topic which is trending on twitter from politics or sport.

- **CREATION OF TWITTER APP:** We already made our twitter app and right now we got our keys. To make Twitter app we followed the following procedure.

  We will click on create app option and fill the essentials



Application Management

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.

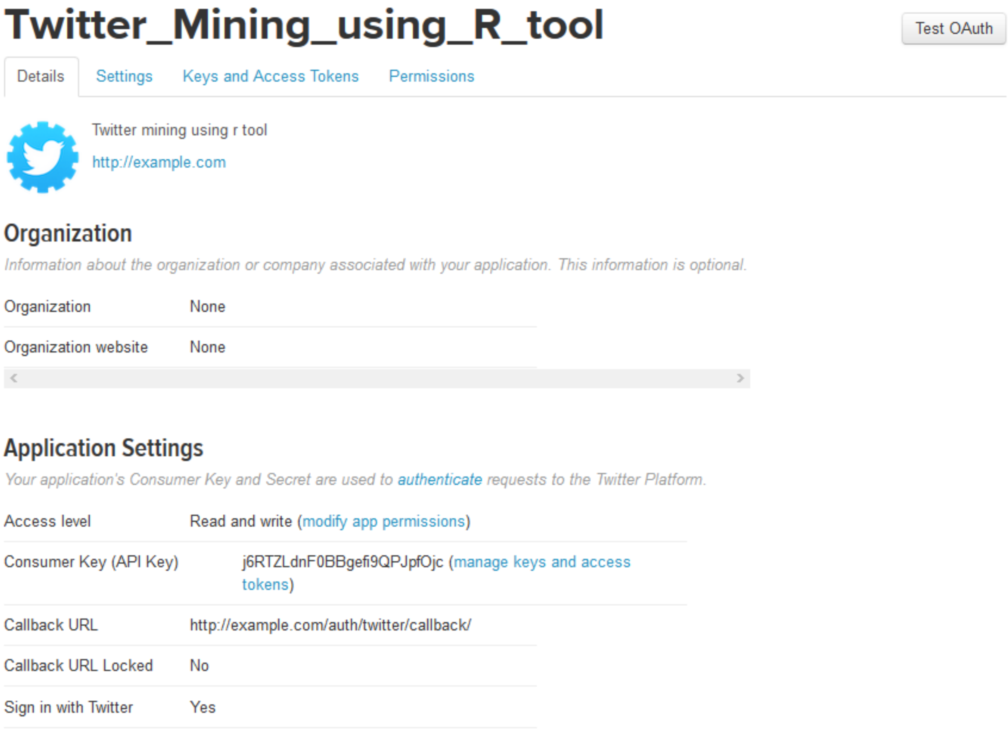
Website *

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-
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 URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token ste
your application from using callbacks, leave this field blank.

Then after the app is created:



Twitter_Mining_using_R_tool

Twitter mining using r tool

http://example.com

Organization

Information about the organization or company associated with your application. This information is optional.

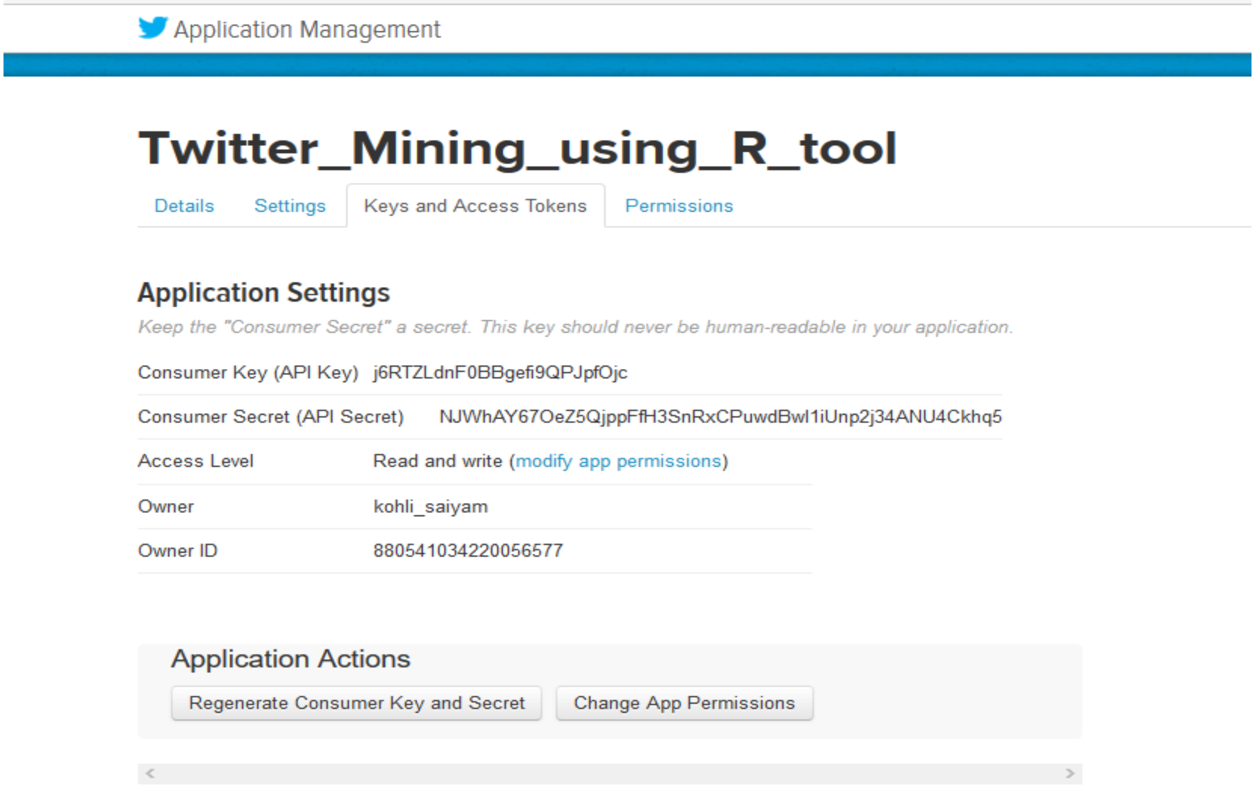
Organization	None
Organization website	None

Application Settings

Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.

Access level	Read and write (modify app permissions)
Consumer Key (API Key)	j6RTZLdnF0BBgefi9QPJpfOjc (manage keys and access tokens)
Callback URL	http://example.com/auth/twitter/callback/
Callback URL Locked	No
Sign in with Twitter	Yes

These are our Access and secret token key which we use for connection with twitter:



Twitter_Mining_using_R_tool

Application Settings

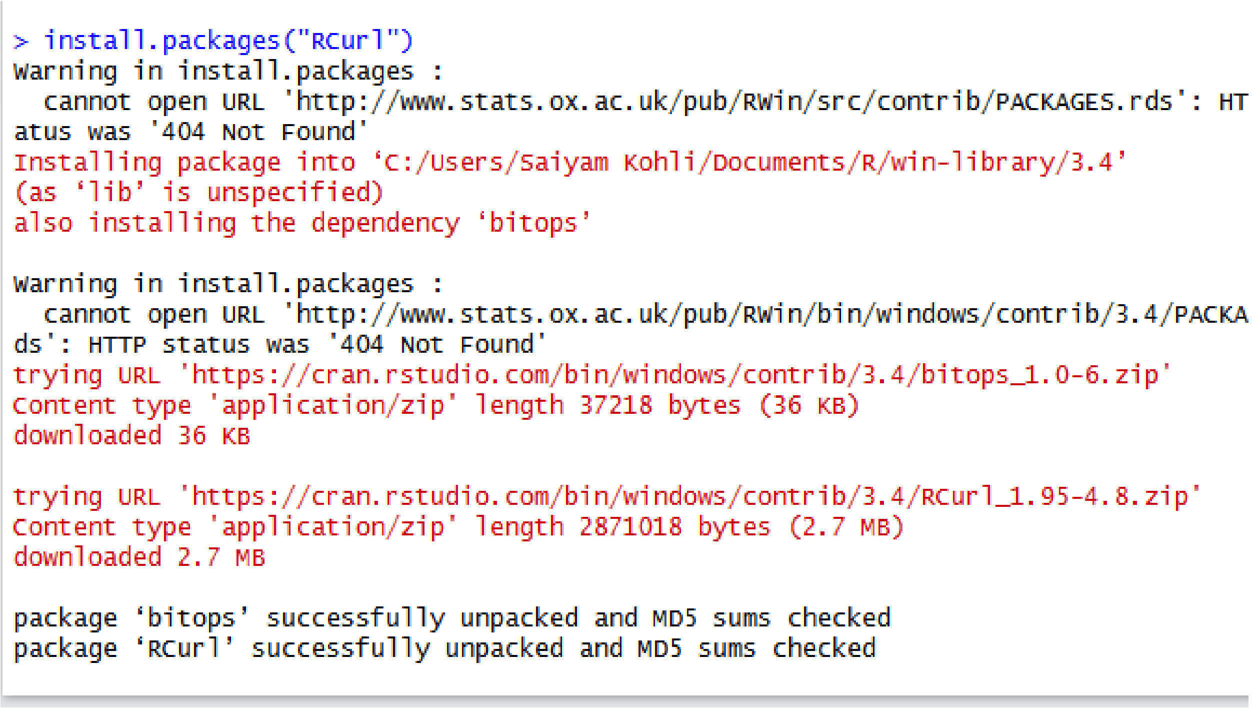
Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	j6RTZLdnF0BBgefi9QPJpfOjc
Consumer Secret (API Secret)	NJWhAY67OeZ5QjppFfH3SnRxCPuwdBwl1iUnp2j34ANU4Ckhq5
Access Level	Read and write (modify app permissions)
Owner	kohli_saiyam
Owner ID	880541034220056577

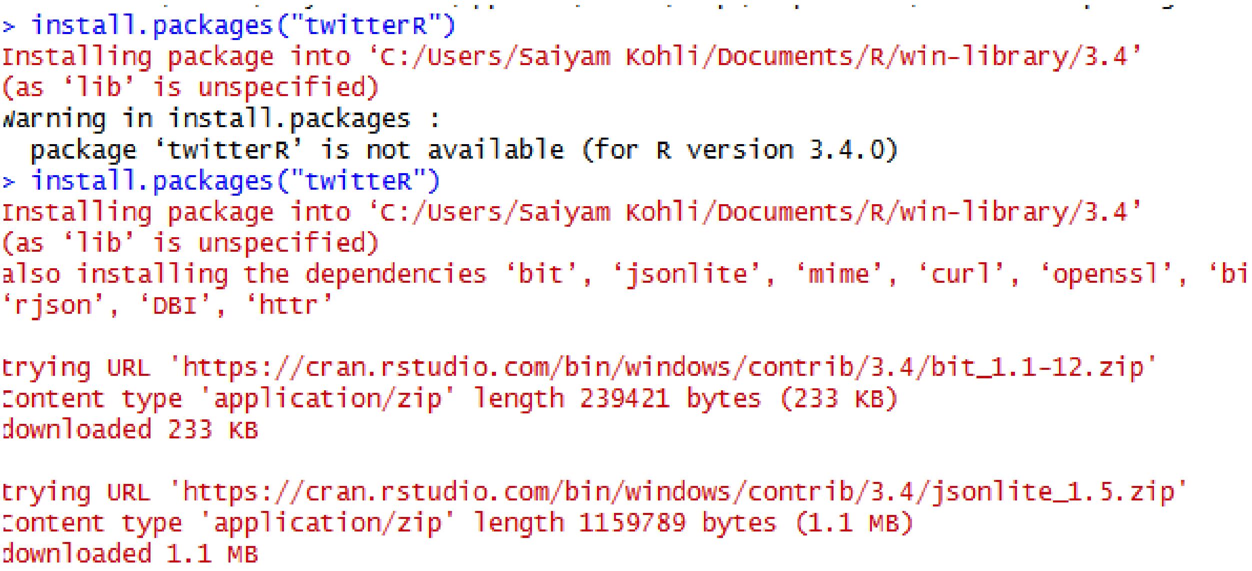
Application Actions

INSTALLING R LIBRARIES AND PACKAGES FOR TWITTER MINING:

For R Part we have to install R Libraries RCURL, TM and WordCloud

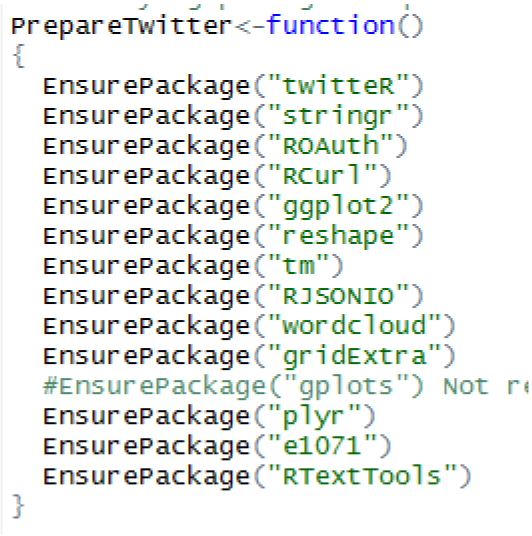


```
> install.packages("RCurl")
Warning in install.packages :
  cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/src/contrib/PACKAGES.rds': HT
atus was '404 Not Found'
Installing package into ‘C:/Users/Saiyam Kohli/Documents/R/win-library/3.4’
(as ‘lib’ is unspecified)
also installing the dependency ‘bitops’

Warning in install.packages :
  cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.4/PACKA
ds': HTTP status was '404 Not Found'
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/bitops_1.0-6.zip'
Content type 'application/zip' length 37218 bytes (36 KB)
downloaded 36 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/RCurl_1.95-4.8.zip'
Content type 'application/zip' length 2871018 bytes (2.7 MB)
downloaded 2.7 MB

package ‘bitops’ successfully unpacked and MD5 sums checked
package ‘RCurl’ successfully unpacked and MD5 sums checked
```



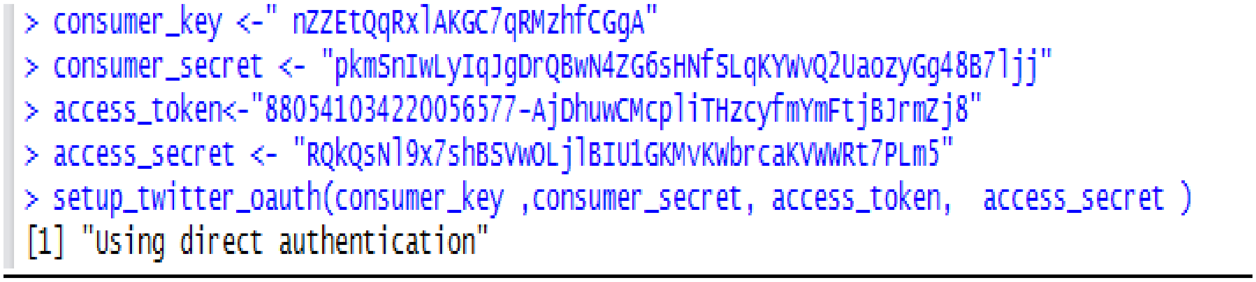
```
> install.packages("twitterR")
Installing package into ‘C:/Users/Saiyam Kohli/Documents/R/win-library/3.4’
(as ‘lib’ is unspecified)
Warning in install.packages :
  package ‘twitterR’ is not available (for R version 3.4.0)
> install.packages("twitteR")
Installing package into ‘C:/Users/Saiyam Kohli/Documents/R/win-library/3.4’
(as ‘lib’ is unspecified)
also installing the dependencies ‘bit’, ‘jsonlite’, ‘mime’, ‘curl’, ‘openssl’, ‘bi
‘rjson’, ‘DBI’, ‘httr’

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/bit_1.1-12.zip'
Content type 'application/zip' length 239421 bytes (233 KB)
downloaded 233 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/jsonlite_1.5.zip'
Content type 'application/zip' length 1159789 bytes (1.1 MB)
downloaded 1.1 MB
```

<u>Packages needed to be Installed</u>

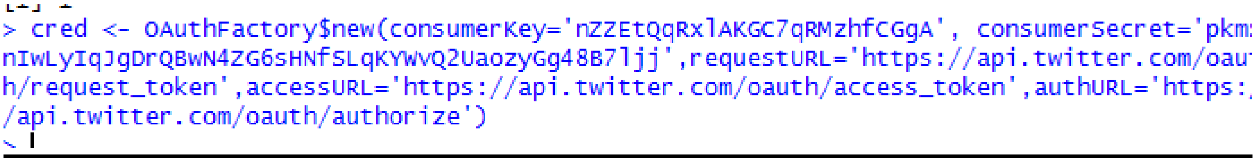


```
PrepareTwitter<-function()
{
  EnsurePackage("twitteR")
  EnsurePackage("stringr")
  EnsurePackage("ROAuth")
  EnsurePackage("RCurl")
  EnsurePackage("ggplot2")
  EnsurePackage("reshape")
  EnsurePackage("tm")
  EnsurePackage("RJSONIO")
  EnsurePackage("wordcloud")
  EnsurePackage("gridExtra")
  #EnsurePackage("gplots") Not re
  EnsurePackage("plyr")
  EnsurePackage("e1071")
  EnsurePackage("RTextTools")
}
```

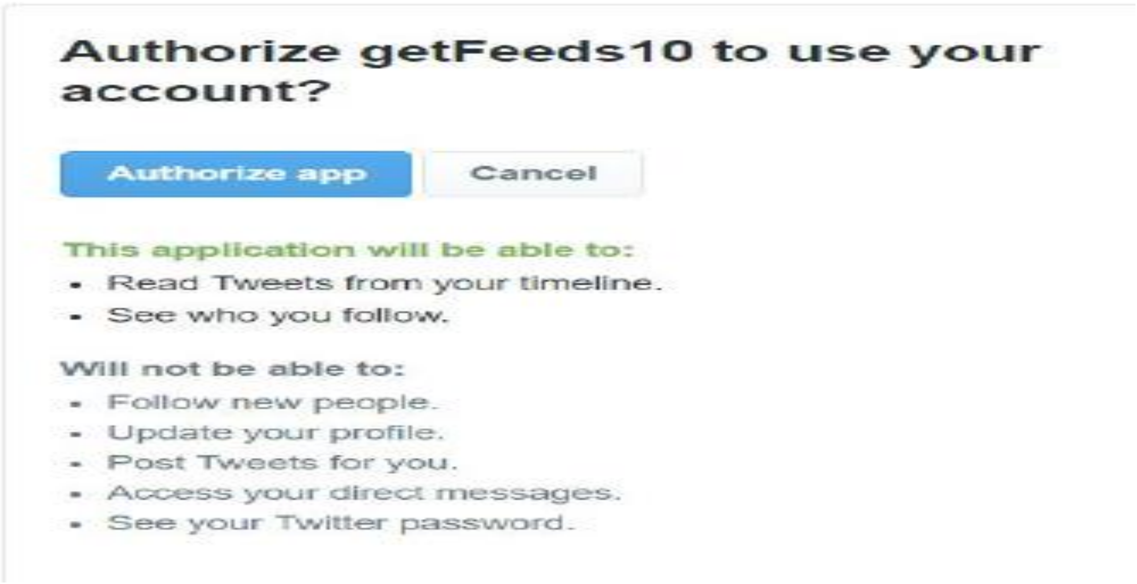
Setting Connection in R with Twitter



```
> consumer_key <-" nZZEtQqRxlAKGC7qRMzhfCGgA"
> consumer_secret <- "pkmSnIwLyIqJgDrQBwN4ZG6sHNfSLqKYWvQ2UaozyGg48B7ljj"
> access_token<-"880541034220056577-AjDhuwCMcpliTHzcyfmYmFtjBJrmZj8"
> access_secret <- "RQkQsNl9x7shBSVwOLjlBIU1GKMvKWbrcaKVWWRt7PLm5"
> setup_twitter_oauth(consumer_key ,consumer_secret, access_token,  access_secret )
[1] "Using direct authentication"
```



```
> cred <- OAuthFactory$new(consumerKey='nZZEtQqRxlAKGC7qRMzhfCGgA', consumerSecret='pkm
nIwLyIqJgDrQBwN4ZG6sHNfSLqKYWvQ2UaozyGg48B7ljj',requestURL='https://api.twitter.com/oau
h/request_token',accessURL='https://api.twitter.com/oauth/access_token',authURL='https:
/api.twitter.com/oauth/authorize')
>
```



Authorize getFeeds10 to use your account?

This application will be able to:

- Read Tweets from your timeline.
- See who you follow.

Will not be able to:

- Follow new people.
- Update your profile.
- Post Tweets for you.
- Access your direct messages.
- See your Twitter password.

Case study: FC Barcelona vs Real Madrid

The tweets are cleaned in R by removing:

● Extra punctuation

● Stop words (Most commonly used words in a language like *the*, *is*, *at*, *which*, and *on*.)

● Redundant Blank spaces

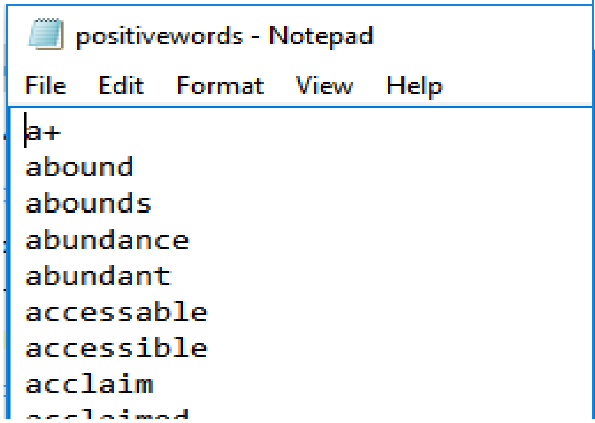
● Emoticons

● URLS

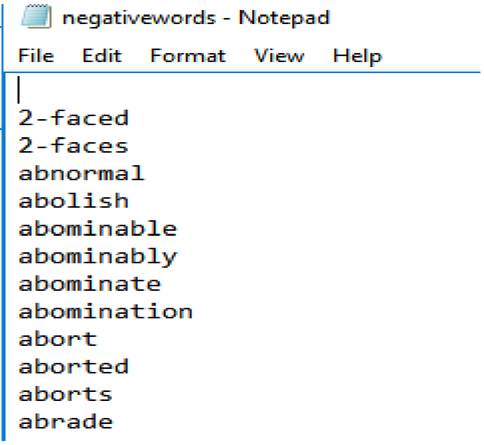
CODE IN R FOR CLEANING(USING GSUB)

- `sentence = gsub('[[:punct:]]',' ',sentence)`
- `sentence = gsub('[[:cntrl:]]','',sentence)`
- `sentence = gsub('\\d+','',sentence)`
- `sentence = gsub('\n','',sentence)`
- `sentence = tolower(sentence)`
- `word.list = str_split(sentence, '\\s+')`
- `words = unlist(word.list)`

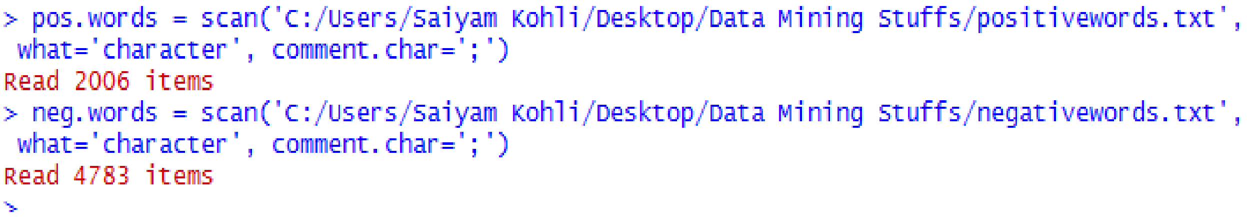
LOADING POSITIVE WORDS AND NEGATIVE WORDS IN DATABASE.



```
a+
abound
abounds
abundance
abundant
accessable
accessible
acclaim
acclaimed
```

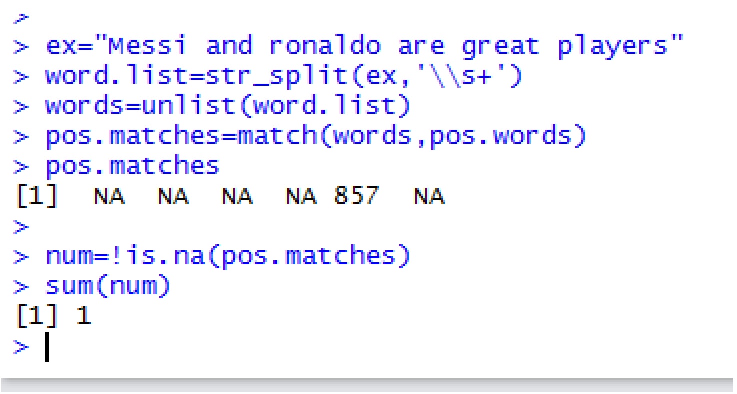


```

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

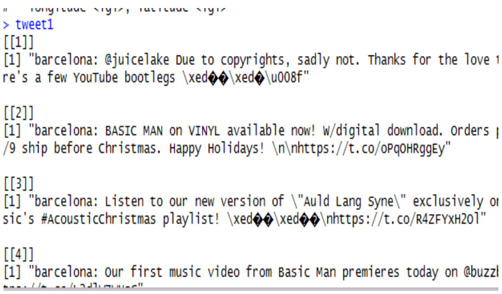


```
> pos.words = scan('C:/Users/Saiyam Kohli/Desktop/Data Mining Stuffs/positivewords.txt',
 what='character', comment.char=';')
Read 2006 items
> neg.words = scan('C:/Users/Saiyam Kohli/Desktop/Data Mining Stuffs/negativewords.txt',
 what='character', comment.char=';')
Read 4783 items
>
```

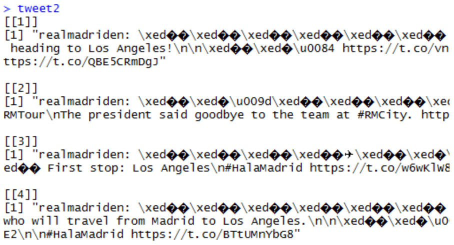
How sentimental analysis works.



```
>
> ex="Messi and ronaldo are great players"
> word.list=str_split(ex,'\\s+')
> words=unlist(word.list)
> pos.matches=match(words,pos.words)
> pos.matches
[1]  NA  NA  NA  NA 857  NA
>
> num=!is.na(pos.matches)
> sum(num)
[1] 1
> |
```

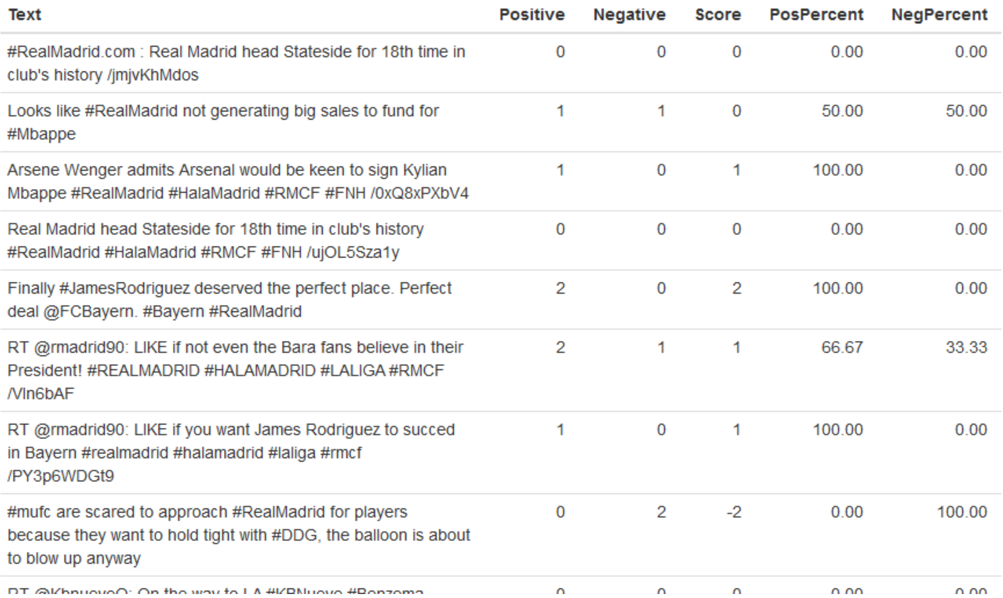


```
#   longitude <lgl>, latitude <lgl>
> tweet1
[[1]]
[1] "barcelona: @juicelake Due to copyrights, sadly not. Thanks for the love t
re's a few YouTube bootlegs \xed��\xed�\u008f"

[[2]]
[1] "barcelona: BASIC MAN on VINYL available now! W/digital download. Orders p
/9 ship before Christmas. Happy Holidays! \n\nhttps://t.co/oPqOHRggEy"

[[3]]
[1] "barcelona: Listen to our new version of \"Auld Lang Syne\" exclusively on
sic's #AcousticChristmas playlist! \xed��\xed��\nhttps://t.co/R4ZFYxH2Ol"

[[4]]
[1] "barcelona: Our first music video from Basic Man premieres today on @buzzl
```



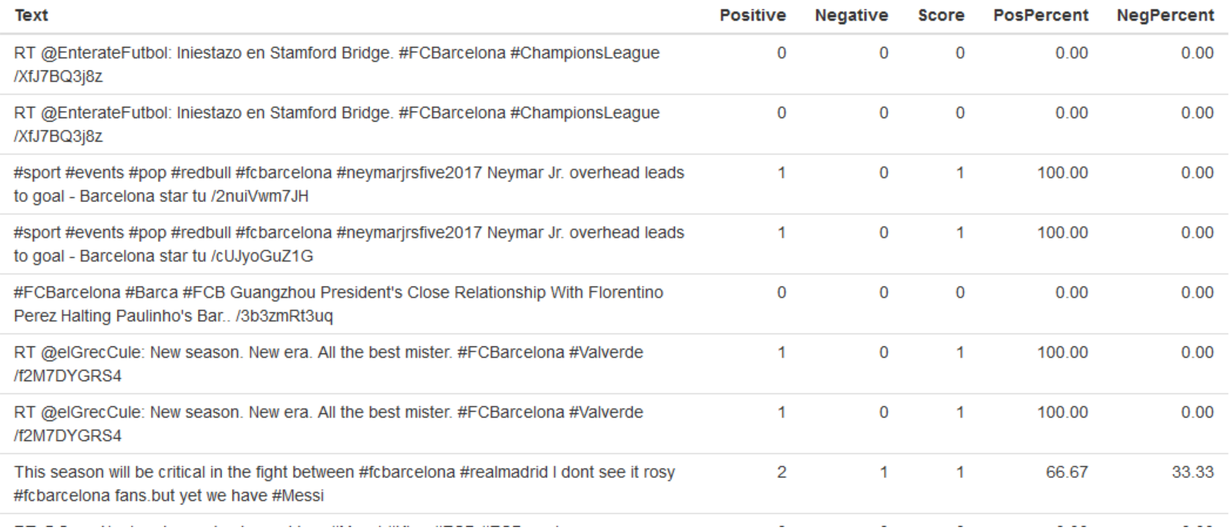
```
> tweet2
[[1]]
[1] "realmadriden: \xed��\xed��\xed��\xed��\xed��\xed��
 heading to Los Angeles!\n\n\xed��\xed�\u0084 https://t.co/vn
ttps://t.co/QBE5CRmDgJ"

[[2]]
[1] "realmadriden: \xed��\xed�\u009d\xed��\xed��\xed��\xed
RMTour\nThe president said goodbye to the team at #RMCity. http

[[3]]
[1] "realmadriden: \xed��\xed��\xed��\xed��✈\xed��\xed�\
ed�� First stop: Los Angeles\n#HalaMadrid https://t.co/w6wKlW8

[[4]]
[1] "realmadriden: \xed��\xed��\xed��\xed��\xed��\xed��
who will travel from Madrid to Los Angeles.\n\n\xed��\xed�\u0
E2\n\n#HalaMadrid https://t.co/BTtUMnYbG8"
```

SENTIMENTAL ANALYSIS FOR REAL MADRID



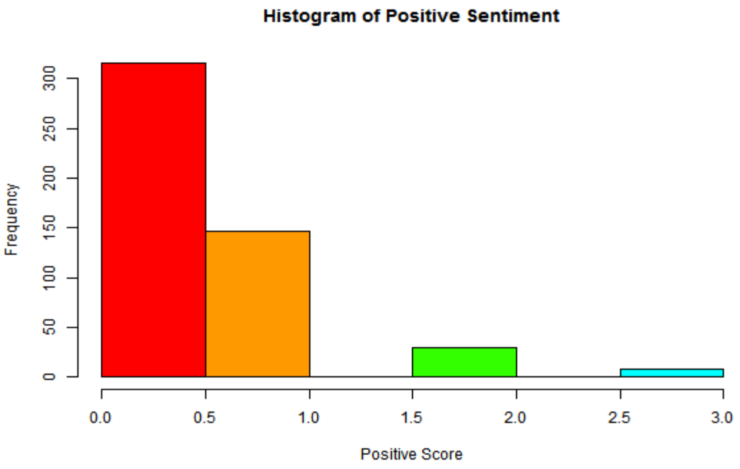
Text	Positive	Negative	Score	PosPercent	NegPercent
#RealMadrid.com : Real Madrid head Stateside for 18th time in club's history /jmjvKhMdos	0	0	0	0.00	0.00
Looks like #RealMadrid not generating big sales to fund for #Mbappe	1	1	0	50.00	50.00
Arsene Wenger admits Arsenal would be keen to sign Kylian Mbappe #RealMadrid #HalaMadrid #RMCF #FNH /0xQ8xPXbV4	1	0	1	100.00	0.00
Real Madrid head Stateside for 18th time in club's history #RealMadrid #HalaMadrid #RMCF #FNH /ujOL5Sza1y	0	0	0	0.00	0.00
Finally #JamesRodriguez deserved the perfect place. Perfect deal @FCBayern. #Bayern #RealMadrid	2	0	2	100.00	0.00
RT @rmadrid90: LIKE if not even the Bara fans believe in their President! #REALMADRID #HALAMADRID #LALIGA #RMCF /Vln6bAF	2	1	1	66.67	33.33
RT @rmadrid90: LIKE if you want James Rodriguez to succed in Bayern #realmadrid #halamadrid #laliga #rmcf /PY3p6WDGt9	1	0	1	100.00	0.00
#mufc are scared to approach #RealMadrid for players because they want to hold tight with #DDG, the balloon is about to blow up anyway	0	2	-2	0.00	100.00
RT @KbnuevoO: On the way to LA #KBNuevo #Benzema	0	0	0	0.00	0.00

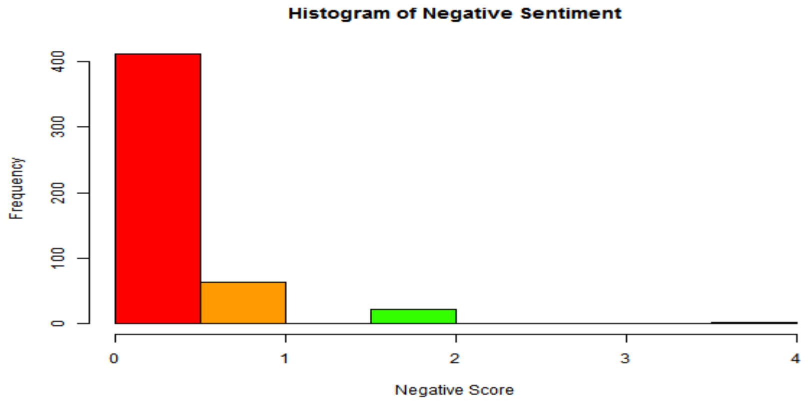
SENTIMENTAL ANALYSIS FOR FC BARCELONA

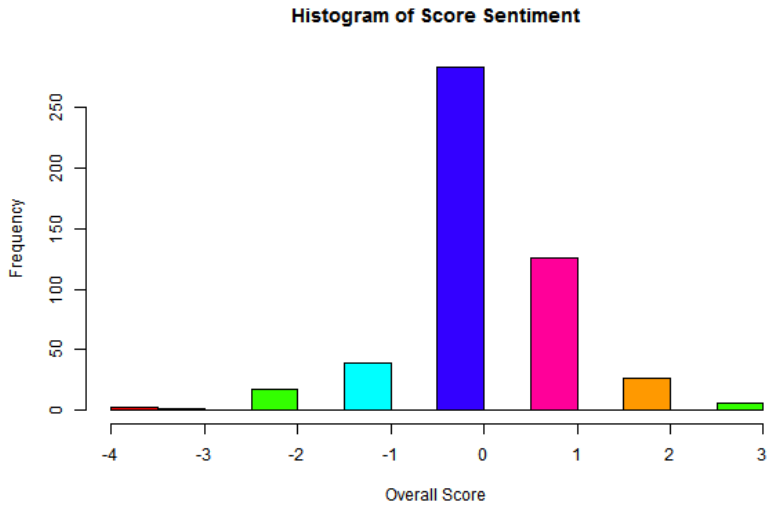


Text	Positive	Negative	Score	PosPercent	NegPercent
RT @EnterateFutbol: Iniestazo en Stamford Bridge. #FCBarcelona #ChampionsLeague /XfJ7BQ3j8z	0	0	0	0.00	0.00
RT @EnterateFutbol: Iniestazo en Stamford Bridge. #FCBarcelona #ChampionsLeague /XfJ7BQ3j8z	0	0	0	0.00	0.00
#sport #events #pop #redbull #fcbarcelona #neymarjrsfive2017 Neymar Jr. overhead leads to goal - Barcelona star tu /2nuiVwm7JH	1	0	1	100.00	0.00
#sport #events #pop #redbull #fcbarcelona #neymarjrsfive2017 Neymar Jr. overhead leads to goal - Barcelona star tu /cUJyoGuZ1G	1	0	1	100.00	0.00
#FCBarcelona #Barca #FCB Guangzhou President's Close Relationship With Florentino Perez Halting Paulinho's Bar.. /3b3zmRt3uq	0	0	0	0.00	0.00
RT @elGrecCule: New season. New era. All the best mister. #FCBarcelona #Valverde /f2M7DYGRS4	1	0	1	100.00	0.00
RT @elGrecCule: New season. New era. All the best mister. #FCBarcelona #Valverde /f2M7DYGRS4	1	0	1	100.00	0.00
This season will be critical in the fight between #fcbarcelona #realmadrid I dont see it rosy #fcbarcelona fans.but yet we have #Messi	2	1	1	66.67	33.33
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

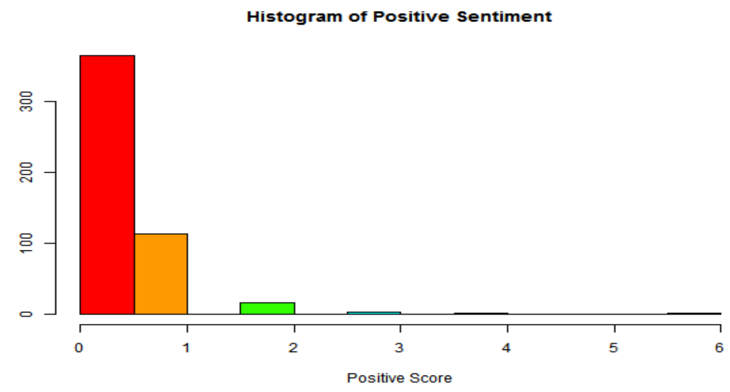
HISTOGRAM FOR FC BARCELONA SENTIMENTS

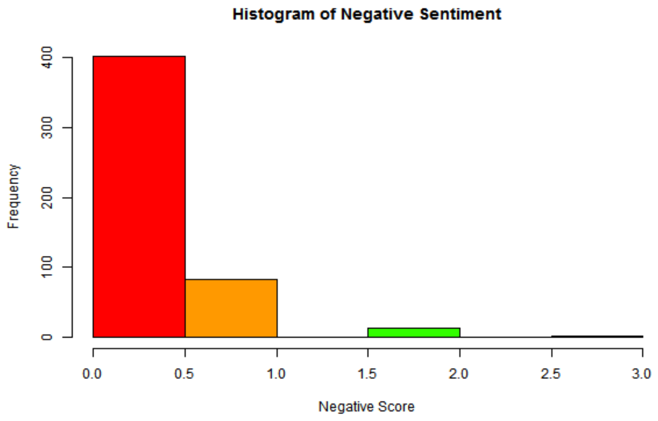


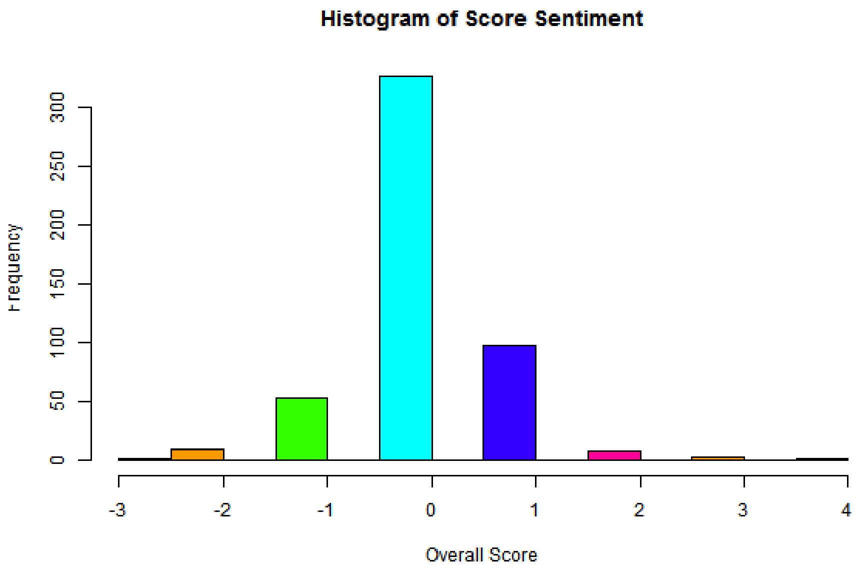




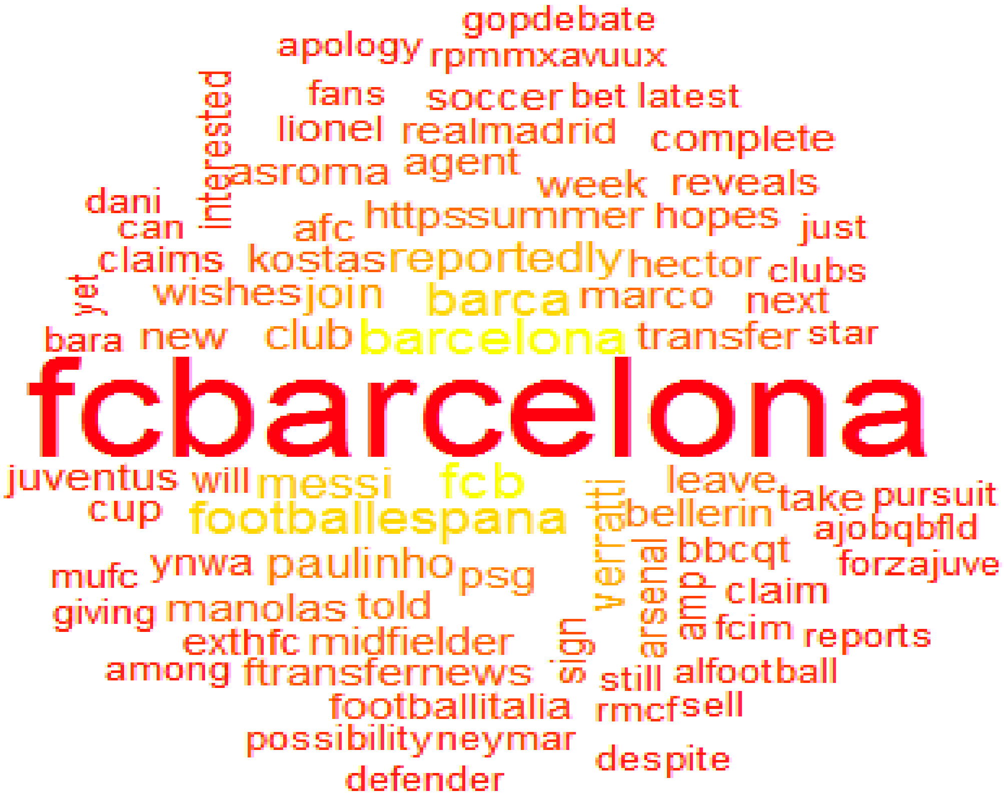
HISTOGRAM FOR REAL MADRID



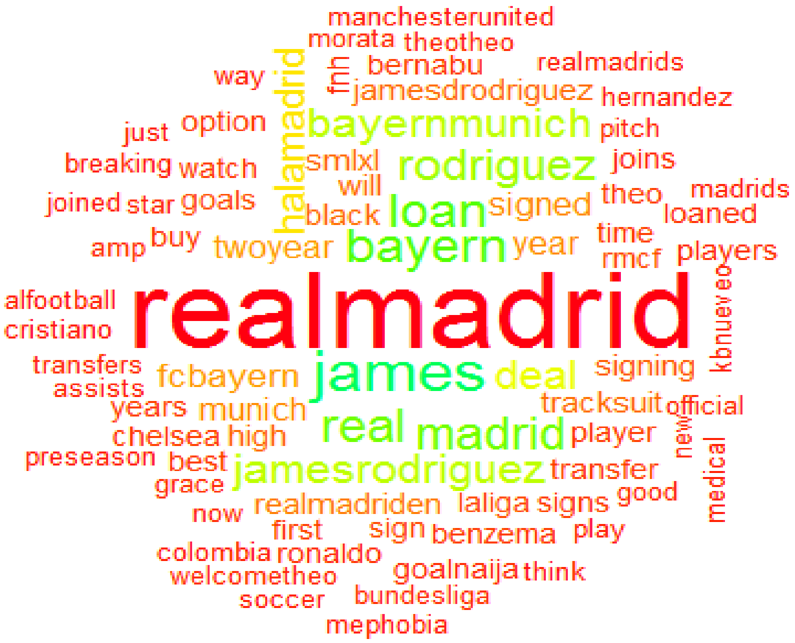




WORD CLOUD FOR FC BARCELONA

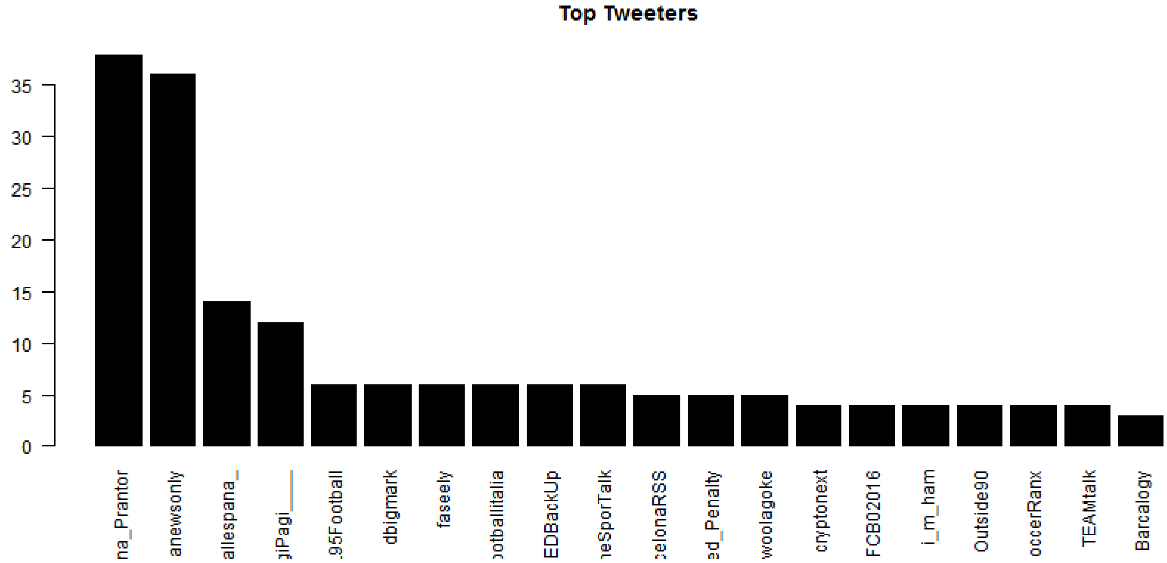


WORD CLOUD FOR REAL MADRID



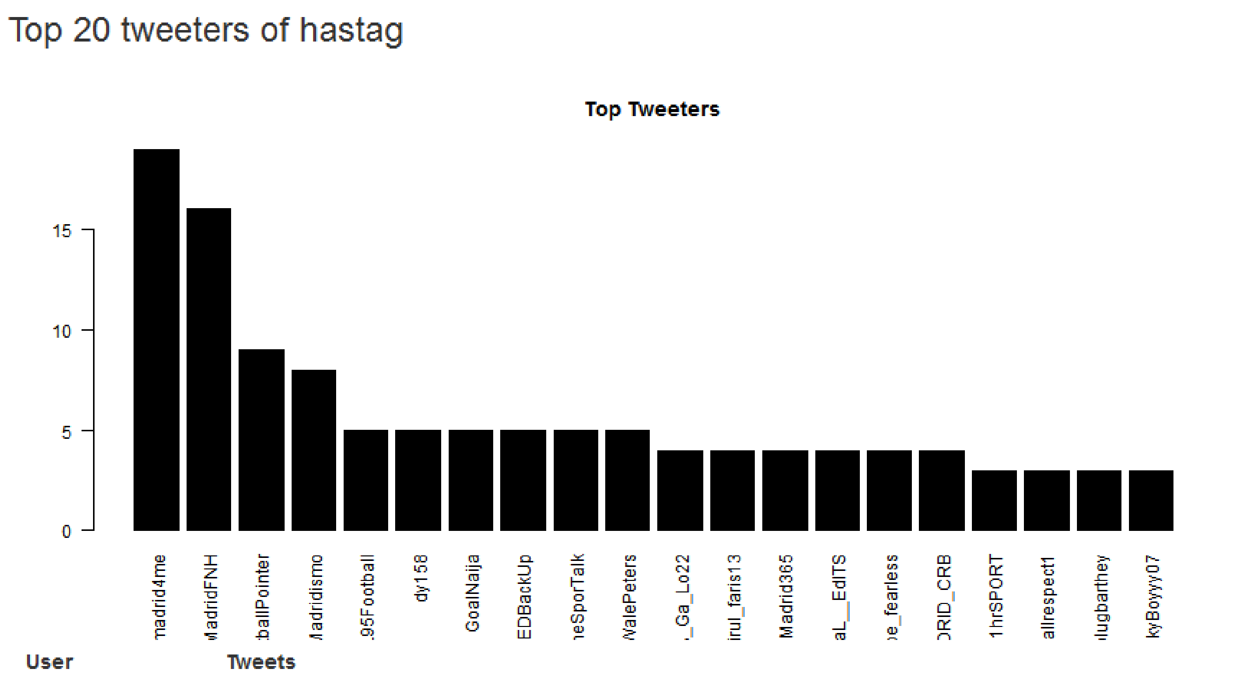
TOP 20 TWITTER HASHTAG FOR FC BARCELONA

Top 20 tweeters of hastag



User Tweets

TOP 20 TWITTER HASHTAG FOR REAL MADRID



Algorithms Used

- **Lexical Analysis:** By comparing uni-grams to the pre-loaded word database, the tweet is assigned sentiment score - positive, negative or neutral and overall score is calculated.
- **Naive Bayes Machine Learning Algorithm:** Training data sets are used to teach the machine what kind of sentences are categorized as positive and what kind are categorized as negative. On arrival of a new tweet or sentence, the machine uses this algorithm to give the correct category to the new data and adds level to the emotion.

PACKAGES USED

- **twitteR** : Provides an interface to the Twitter web API
- **stringr** : String operations in R
- **ROAuth** : Provides an interface to the OAuth 1.0 specification allowing users

to authenticate via OAuth to the server of their choice.

- **RCurl** : Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server.

- **ggplot2** : An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources.

- **reshape** : Flexibly restructure and aggregate data using just two functions: melt and cast

- **tm** : A framework for text mining applications within R.

- **RJSONIO** : This is a package that allows conversion to and from data in Javascript object notation (JSON) format. This allows R objects to be inserted into Javascript/ECMAScript/ActionScript code and allows R programmers to read and convert JSON content to R objects

- **wordcloud** : visual representation in the form of wordcloud where size of the word is proportional to the frequency of words used in the tweets

- **gridExtra** : Provides a number of user-level functions to work with "grid" graphics, notably to arrange multiple grid-based plots on a page, and draw tables.

- **plyr** : Tools for Splitting, Applying and Combining Data

LIMITATIONS

1. The Twitter Search API can get tweets upto a maximum of 7 days old.
2. Not effective in detecting sarcasm.
3. Cannot get 100% efficiency in analysing sentiment of tweets.
4. Can only retrieve a maximum of 1000 tweets per query without authenticating via OAuth before receiving a 403 error or timeout.
5. Giving a hash tag under the wrong category will still give results: No error message