

National College of Ireland
BSc. (Hons) in Computing – Data Analytics
2016/2017

Fasial Bashar

X13358851

x13358851@student.ncirl.ie

fasialbd2010@yahoo.ie

EPL Analysis: Sentiment and Predictive Analysis
Technical Report



Declaration

SECTION 1 *Student to complete*

Name: FASIAL BASHAR
Student ID: X13358851
Supervisor: MUHAMMAD IQBAL

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: _____

Date: _____

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to you being suspended or expelled from college.

Complete the sections above and attach it to the front of one of the copies of your assignment.

Table of Contents

Executive Summary	5
1 Introduction.....	6
1.1 Background	6
1.2 Aims	7
1.3 Technologies.....	7
2 System	8
2.1 Requirements	8
2.1.1 Functional requirements	8
1.1.1 Requirement 1: Setup API to gather Data	8
1.1.2 Requirement 2: Clean Data	10
1.1.3 Requirement 3: Data Classification.....	12
1.1.4 Requirement 4: Analyse Data	13
1.1.5 Requirement 5: Output the Result.....	15
2.1.2 User requirements	16
2.1.3 Non-Functional Requirements.....	17
2.1.4 Environmental requirements	18
2.1.5 Usability requirements.....	18
2.1.6 Data requirements	18
2.2 Analyse and Design	19
2.3 Methodology:.....	20
2.4 Machine Learning Algorithms	21
2.4.1 Score.Sentiment:.....	21
2.4.2 Naïve Bayes:	22
2.5 Implementation	23
2.5.1 Data Mining	23
2.5.2 Data Analysis:	28
2.6 Testing and Evaluation	34
3 Conclusions	41
4 Further development or research.....	42
5 References	43

6	Appendix: A.....	45
6.1	Project Proposal	45
6.2	Project Plan	49
6.3	Monthly Journals	50
6.4	Appendix B: Python Script for Data Mining	57
6.5	Appendix C: Manual Classification.....	58
6.5.1	MCILIV	58
6.5.2	WBAARS	62

Executive Summary

Sentiment analysis is also known as opinion mining, is a machine learning method to extract sentiment from text and databases. Sentiment analysis is fast growing method used by many companies in many sectors of business to help them understand voice of people based on their online reviews or comments on social media like Facebook or Twitter. Sentiment analysis focuses to determine the attitudes, emotions and opinion of a person based on their text or document. Sentimental analysis algorithms can group the text or tweets based on the opinion, attitude and emotions.

Twitter is one of the major social media services, where people share their voice or opinions about everything. Football is one of the most popular topics on twitter where people share their opinion, from a tweet being about their favourite team or a rival team. Everyone has an opinion, which they like to share with the world, they can use twitter handle (@ManUtd) to get the message across the team or person they are talking about which people can add on their tweets.

The topic this project will focus on is Football and more specifically the English Premier League. Live Tweets will be gathered for matches involving multiple teams. Then the data will be analysed and machine learning algorithm will be used to score “Positive” or “Negative” for each tweet. All the score will be displayed through the help of visualization.

The main objective of the project is to find out how fans react during the premier league matches by collecting the tweets during the matches and running sentimental analyses on the tweets.

1 Introduction

1.1 Background

The reason for choosing this project was simple, I thought I should project on something I like and will enjoy doing. Football is a popular sport and most followed sports in the world. Wherever people are in the world, they know about and they follow football. Everyone has that favourite team that they support in their country or some other team in different countries. The main reason for choosing football to analyse is that I wanted to know how people react to a certain team or match. Although there are many top football leagues around the world, I decided to focus on the English Premier League (EPL). Fans everywhere want their opinions to be heard from other fans and everyone has certain views/opinions on certain matches or certain teams. Social media's like Facebook and Twitter are used by most fans to show their disappointment or excitement after a match of their favourite team that won or lost. Twitter doesn't categorise the tweets as positive or negative, I so thought maybe I can show it through my project.

The aim of this project is to showcase the sentiment analysis with the help of visualisation and can compare if the fans' opinions have changed. How fans are feeling towards the games based on the result from machine learning algorithms.

Sentiment analysis is also known as opinion mining, identifying and categorising the text is the main objective. The data can be expressed in text form, to determine the user's attitude towards certain topics. Sentimental analysis has become very popular in the marketing area, where a certain organisation wants to know positive and negative things people say towards them. There is some powerful data mining software that is available to data scientist.

1.2 Aims

The aim of the project is to develop a model based on data analytics. The purpose of the project is to perform sentimental analyse on English Premier league and perform a prediction algorithm on the data that will be gathered throughout the project. R and Python will be used to perform sentiment analysis and will be used to show the result of the analysis. The sentiment analysis will classify the output result as Positive, Neutral or Negative. During the match day, tweets will be gathered and saved to csv files. Tweets will be gathered for each match and saved separately. Premier league uses a hashtag for each match so that the fans can tweet specifically for that certain match. For example, if Arsenal plays Manchester United the hashtag will be #MUNARS, like that there is a hashtag for every match during the premier league season. Which will me to gather the data that is related this project. Another purpose of this project is to compare the result from multiple machine learning algorithm and see which is more accurate when it comes to categorising based on polarity and sentiment.

1.3 Technologies

The majority work on this project will be completed using R. R is a language and environment for statistical computing and graphics. "It is a GNU project which is like the S language and environment which was developed at Bell Laboratories by John Chambers and colleagues". R provides a wide range of statistical and graphical techniques, and is highly extensible. R also comes with many inbuilt libraries that can be used to perform analysis.

Other Technologies that might be used for this project:

- Python will be used to write the script that will be used to gather data from Twitter via Streaming API
- Excel will be used to save the data that will be gathered in csv format.
- Notepad++

2 System

2.1 Requirements

2.1.1 Functional requirements

The functional requirements of the system that will be required to complete the project:

1. The user setups streaming API for twitter by writing code in Python or R
Gathers data from twitter via twitter API and save to external database. The user will also acquire Data from websites.
2. The system will cleanse the data
3. The system will classify data (Positive, Natural and Negative)
4. The system will analyse data
5. The result will be shown through visualization

1.1.1 Requirement 1: Setup API to gather Data Description & Priority

To carry out the project, data will be required and to get the live data during match we need to setup streaming API. We can do this by writing a small piece of code, which can be done in R or Python language. Data will also be gathered from an external website.

Use Case

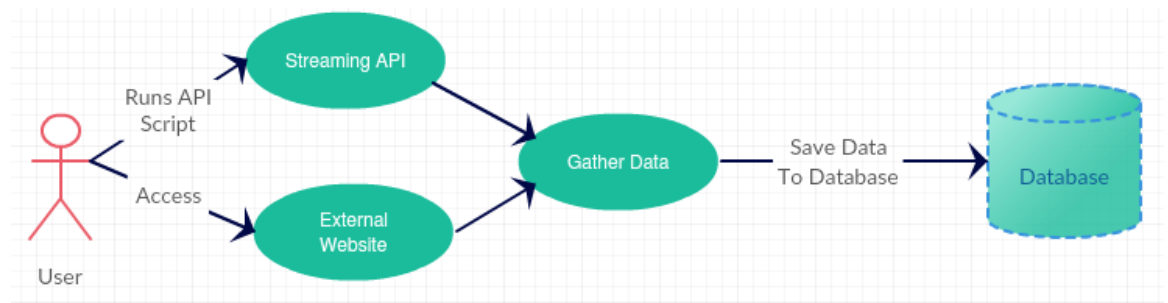
Scope

The scope of this use case is to get data by getting live tweets during the matches and from external resources (Web)

Description

This use case describes the process of the user getting the data that will be required for the system.

Use Case Diagram



Flow Description

Precondition

This is the first step of the project, data collection. When the code is initialized, the system will connect to twitter database and save the data that will be saved to an external database. The user will also access a Football database website and acquire the data that is required.

Activation

This use case begins when the user runs the Python and R script. To get statistical data the user needs to access website and download the data.

Main flow

1. The user initiates the process
2. The User runs the script through Python or R
3. The system connects to the twitter database; it obtains the user credential from the script.
4. Once its connected, API begins to stream the database to an external database.
5. The user access websites and download data
6. Once the data are collected, the user can terminate the process.

Exceptional flow

E1:

1. The system does not initiate
2. The User not able to run the API script or access the website for data

Termination

The user can terminate the data mining process by stopping the programme because the required data have been collected or if it reaches the end of the streaming process from twitter.

Post condition

The project is ready to move on to the next step

1.1.2 Requirement 2: Clean Data

Description & Priority

This part of the requirement focuses on cleaning the data that we have collected from twitter via the API. We will do this by eliminating unwanted and unusable data so that we can get an accurate result.

Use Case

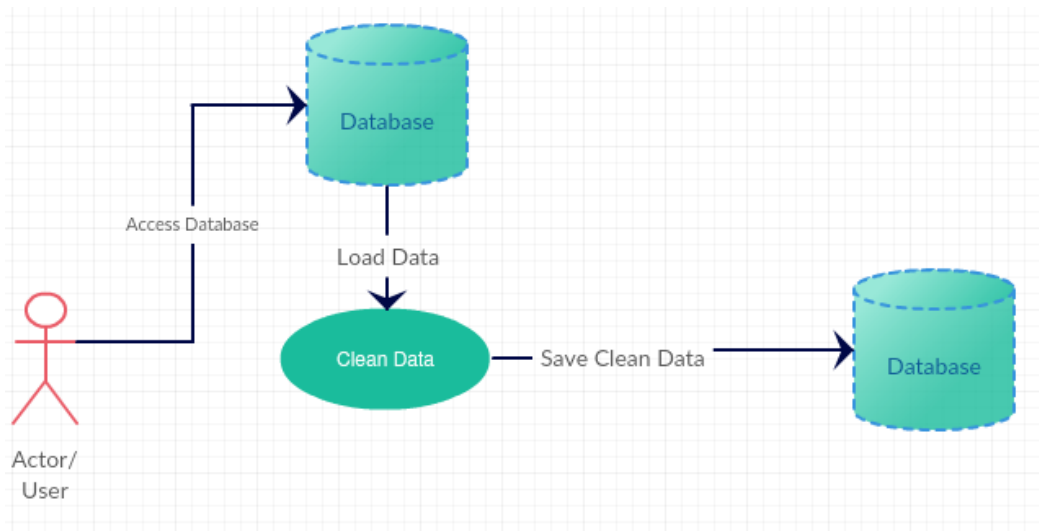
Scope

The scope of this is to clean data by removing unwanted and unusable data.

Description

This use case explains the process of cleaning data to be analysed later.

Use Case Diagram



Flow Description

Precondition

The system is ready to clean the data, after the user loads the data it will be ready to start the data cleaning process.

Activation

This use case begins when the user uses Python or R to clean the collected data.

Main flow

7. The user initiates the process
8. The user reads the data from the database or CSV file to cleaning tool
9. The system removes unwanted data
10. The system replaces the old data with new clean data and saves to the database
11. The system terminates

Exceptional flow

E1:

3. The user couldn't load the data file

Termination

Once the cleaning process is completed, the system will be terminated automatically or stopped manually by the user.

Post condition

The cleaned data will be stored with clear file name. There will not be duplication of any of the files.

1.1.3 Requirement 3: Data Classification

1.1.3.1 Description & Priority

The data classification will be used to categorise the clean data and the classification that will be used are positive, Neutral and Negative. This can be done using R language; this process will be an important step of the project as it can influence the result. Each classification is known as sentiment, which is important for this project as sentiment analysis is the focus of this project.

1.1.3.2 Use Case

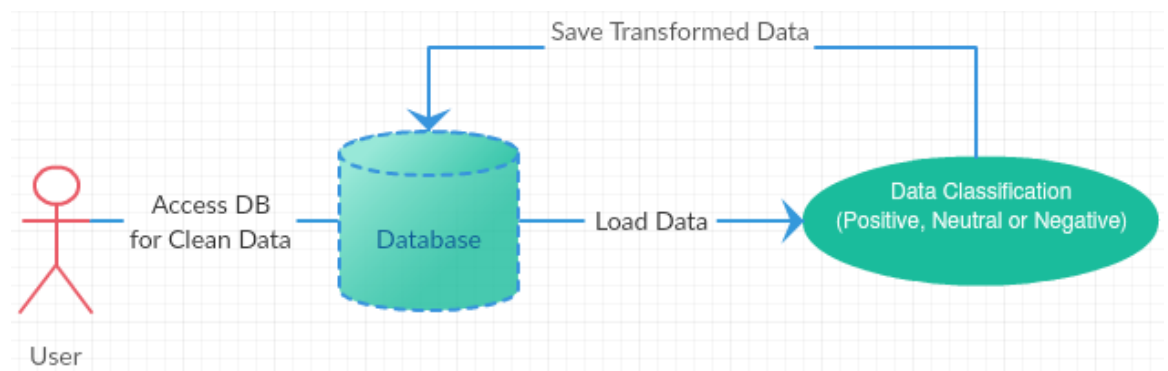
Scope

The scope of this to categorise the data into Positive, Neutral or Negative.

Description

This use case explains the process of the data being classified.

Use Case Diagram



Flow Description

Precondition

The precondition of this step was clean the data by removing unwanted and unusable data and store the clean data afterwards to the database.

Activation

This use case begins when the user loads the data from the database and uses R to check how positive, neutral or negative the data is.

Main flow

12. The user initiates the process
13. The user loads the clean data to the system
14. The system runs sentimental analyses on the data
15. Once the system analyses all the data, system will terminate

Exceptional flow

E1:

4. The data couldn't be classified, as the system gives an error.

Termination

The system will terminate once all the data have analysed.

Post condition

The system is ready to perform the next step of the project.

1.1.4 Requirement 4: Analyse Data

1.1.4.1 Description & Priority

After classifying the data, we can now analysis the data. This stage is very crucial part of the project as the result we get will depend on how the analyse is running.

1.1.4.2 Use Case

Each requirement should be uniquely identified by a sequence number or a meaningful tag of some kind.

Scope

The scope of this use case is to run the final analysis the data and get

Description

This use case describes the process of analysing the data.

Use Case Diagram



Flow Description

Precondition

The system is in initial mode and ready for the data to be loaded and once it's loaded the user can perform their analysis.

Activation

This use case begins when the user uses an R program to analyse the data.

Main flow

16. The system user initiates the process
17. The data being read into the system
18. The system will run the analysis required by the user
19. The system will output the result
20. Once completed, the system will terminate

Termination

The system will terminate after the all the analyses have been performed on the data.

Post condition

The system goes into a waiting mode for the next step of the project.

1.1.5 Requirement 5: Output the Result

1.1.5.1 Description & Priority

This is the final requirement of the project; this is where we get the result of the project. The result will be shown through the help of R programmes built in visualisations.

1.1.5.2 Use Case

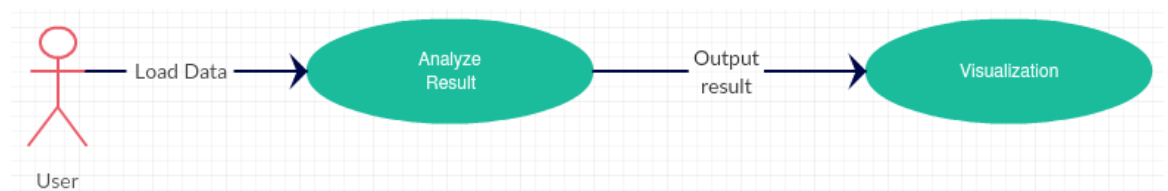
Scope

The scope of this is to output the result of the analysis that was performed in the previous stage. This is the final stage of the project; this stage will be very crucial as this is where the result will be shown via visualisations.

Description

This use case explains the how the system will display the result with the help of R programs built in visualisation.

Use Case Diagram



Flow Description

Precondition

The system is in initial mode and ready for the last and final part of the project, which is to display the result with the help of visualisation.

Activation

This use case begins when the user begins writing scripts in the R studio to analyse the data.

Main flow

21. The user initiates the system

22. The system identifies the data file
23. The user writes code on R
24. The system will output the result in the form of visualization or other
25. The user terminates the system after the task is completed

Exceptional flow

E1:

5. The system couldn't publish the result
6. The user couldn't show their analysis result due to system's faults.

Termination

The system terminates after the user completes the task and achieves a desirable result.

Post condition

The system goes into a waiting mode.

2.1.2 User requirements

The primary objective of the system is to do a sentimental analysis on data that I'll be gathering from twitter during the premier league matches and data will be analysed with multiple machine learning algorithm. Result from the algorithms will be compared.

The completed script should be able to:

- Gather Data using Python and R Script
- Show tweets as categorised by sentiment
- Show Compared result of two algorithms and evaluate the result.
- Compare positive and negative result by visualizing.
- Perform a predictive algorithm on the data

2.1.3 Non-Functional Requirements

Performance/Response time requirement

If the user has Python or R tool, then they just load the script file into it. From there the data should load immediately and output desirable result for the user.

Availability requirement

As this system will not need a connection to the Internet, so there will be no downtime. They will just load in the script from the computer and it should be accessible all time.

Recover requirement

The main copy of the data file and the script file will be stored cloud based website (e.g.: Dropbox or Github), which can be recovered anytime. If the user accidentally deletes the copied file from their computer/Laptop.

Robustness requirement

If one part of the script has an error the system will still, try run the working part of the script. It doesn't depend on just part of the script, as there will be multiple codes for multiple analysis. The system (R Studio) will tell the user which part of the code has an error so that they can resolve that error.

Reliability requirement

The system will be able to run the script and the database anytime the user wants and if the user have a specific tool to run the files as it doesn't need specific time to be used.

Maintainability requirement

If the script or the database has some minor error, it will be easy to correct it. If there is wrong code in the programming script, the user can easily access the file and correct it.

Scalability requirement

The script can get larger and larger as the project goes on, there will not be any problem with scalability as there is no restriction on how big the file should be.

Portability requirement

This project will be portable; the script of code can be saved onto a USB stick or the cloud. The user will be able to carry it around and use whenever and wherever want.

Extendibility requirement

The user can expand the code by continuing to write more code to analyse same database or they can use to analyse much bigger data set. Multiple Machine learning algorithms can be performed on the same dataset.

Reusability requirement

Python script can be used for data mining from twitter based on the keywords. R script can be used to analyse dataset containing different tweets.

2.1.4 Environmental requirements

If the user has R studio available on their computer, then they can run the script and the result will be shown in form of visualisation. R-studio is free to install on any operating system.

2.1.5 Usability requirements

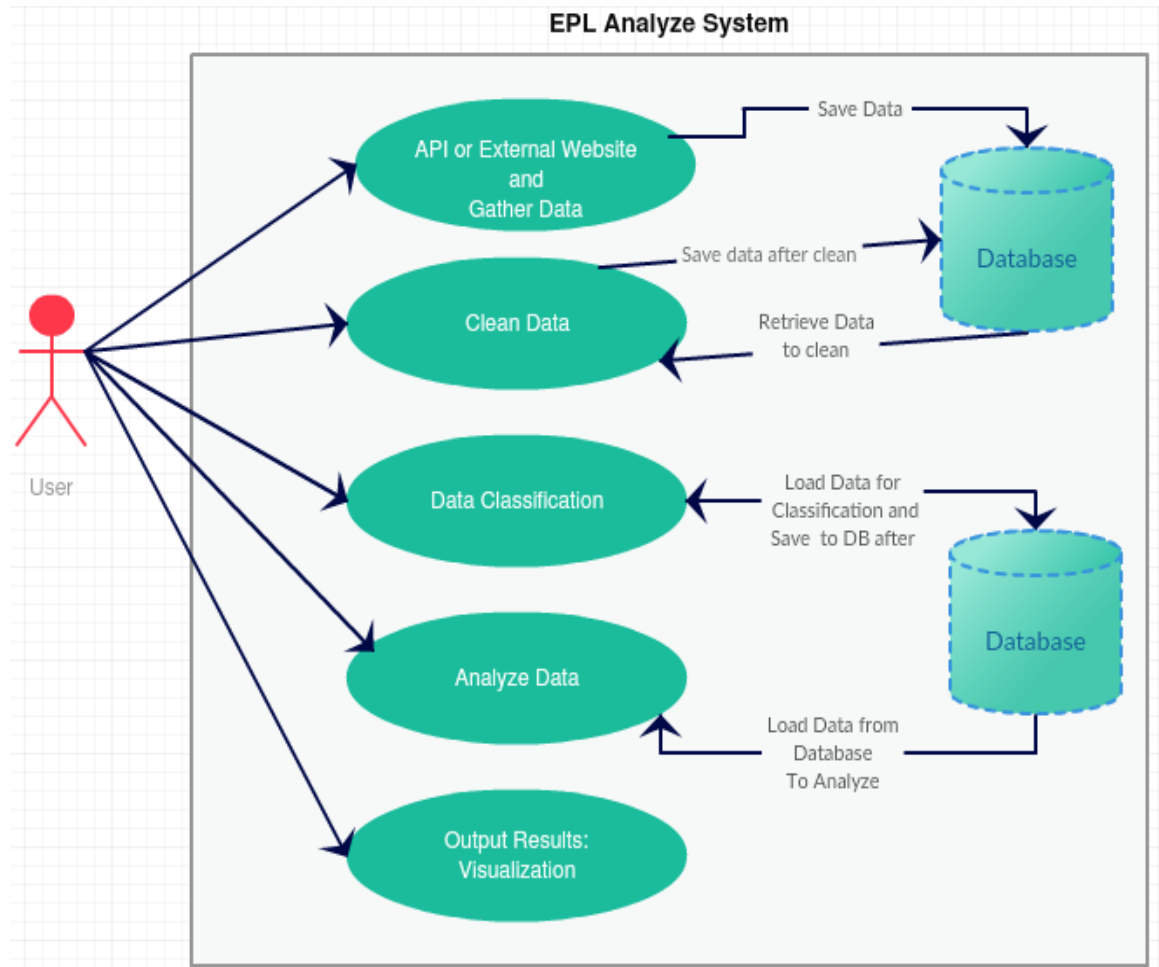
The script must have simple R or Python language so that users can understand or add comments so that will help the user to understand the concept.

2.1.6 Data requirements

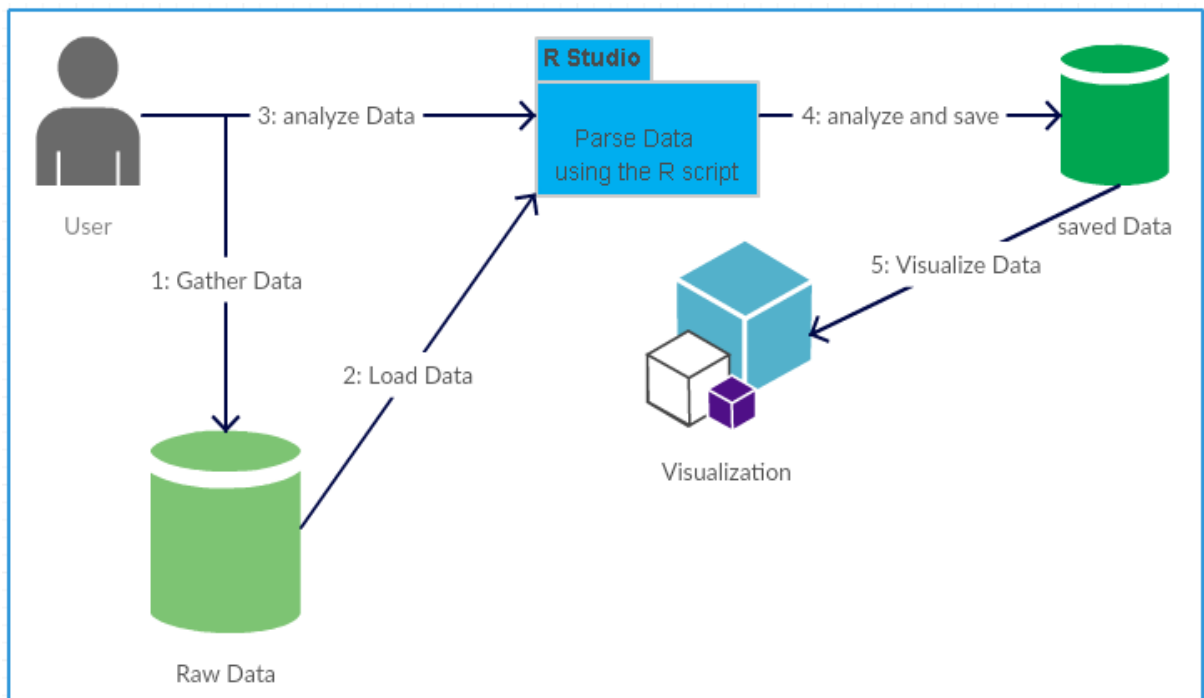
R can read files in multiple formats, e.g. CSV, txt and from MySQL. The user can change the format in the R script so that they can analysis data in any format they desire.

2.2 Analyse and Design

Use Case Diagram:



Logical Architecture:



2.3 Methodology:

There is multiple methodology that could've been used for this project, but the methodology that was used for this project is KDD, also known as Knowledge Discovery and Data Mining. *Knowledge Discovery in Databases (KDD)* refers to the process of finding knowledge from data in large databases. There are a few important steps involved in the KDD, to achieve the desired result. The diagram below shows the steps involved in KDD Process: (Leondes, 2000), (DBD, 2016)

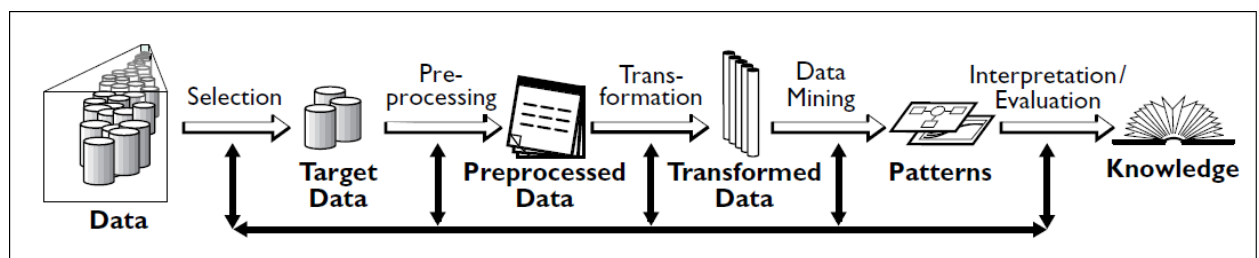


Diagram: KDD Process, step by step.

The KDD process steps outlined:

- **Selection**
This step consists of selecting a target data set or data sample, on which discovery is to be performed. For this project, I have targeted the premier league tweets as my data, I used a streaming API to gather that data.
- **Pre-Processing**
This part consists of the target data being cleaned and pre-processing to obtain consistent data. To clean the twitter data, I removed all the unwanted data and special characters.
- **Transformation**
This stage consists of the data transformation by using dimensionality reduction or transformation methods.
- **Data Mining:**
This stage consists of on the searching for patterns of interesting in a representational form, depending on the data mining objective (usually, prediction)
- **Interpretation/ Evaluation:**
This stage consists of the interpretation and evaluation of the mined patterns. This is also documented for further study or usage.

2.4 Machine Learning Algorithms

The focus of this project is sentiment analysis and predictive analysis. Sentiment analysis focuses to determine the attitudes, emotions and opinion of a person based on their text or document. Sentiment analysis algorithms can group the text or tweets based on the opinion, attitude and emotions. R-studio offers packages that contain functions that allows the users to carry out sentimental analysis.

2.4.1 Score.Sentiment:

Score.senrtiment algorithm is an easy and effective algorithm that assigns scores to each tweet by counting how many “positive” or “negative” or “neutral” words are in the tweet. There is a simple way to calculate score for each tweet: $\text{Score} = \text{Total Positive Words} - \text{Total Negative Words}$.

Positive: When the score of the tweet is greater than 0 ($\text{Score} > 0$), the tweet count as “positive”.

Negative: When the score of the tweet is less than 0 ($\text{Score} < 0$), the tweet count as “negative”

Neutral: When the score of the tweet is equal to 0 ($\text{Score} = 0$), the tweet count as “neutral”

For the score.sentiment algorithm to count how many words are positive or negative, we need to add lexicon of words or also known as word dictionary to R. This dictionary contains 2006 Positive and 4782 Negative words. This dictionary was gathered by Hu and Liu.

2.4.2 Naïve Bayes:

The Naive Bayes algorithm is a machine learning algorithm that is based on probability model. Naïve Bayes is commonly used for text classification in many applications. The algorithm is known as “Naïve” because it assumes that all the attributes related to the dataset are important and independent. Although most of the times assumptions are incorrect, still this algorithm tends to be the first method of choice for classification learning due to its high accuracy with many conditions and versatility. The implementation is simple and doesn’t require high computational power. It’s a fast and effective technique used in many opinion mining applications. The algorithm works by assuming all the words in the dataset are unique and are unrelated to all the words in that dataset. The algorithm doesn’t know the differences between words and sentences, so it assumes everything independent. The Naïve Bayes algorithm doesn’t have any limits of how big or small a training dataset, it works very well with all data. It uses all the feature of a dataset to make a classification, which is a big advantage the algorithm.

For this project, I have utilized the classify_emotion and classify_polarity function that uses the Naïve Bayes algorithm to classifies the tweets by emotion and polarity. These functions are available to R, through the “sentiment” package.

2.4.2.1 *classify_emotion:*

This function allows us to analyse tweets, by labelling each tweet by different types of emotions: surprise, joy, sadness, fear, disgust and anger. This can be done by using two algorithms, one is Naïve Bayes and the other is a simple voter procedure.

2.4.2.2 *classify_polarity:*

The `classify_polarity` is a different compare to the classification of emotion, the `classify_emotion` classifies data by six emotions. Whereas the `classify_polarity` classify text as negative or positive. The classification can be done by using the Naïve Bayes Algorithm.

2.5 Implementation

2.5.1 Data Mining

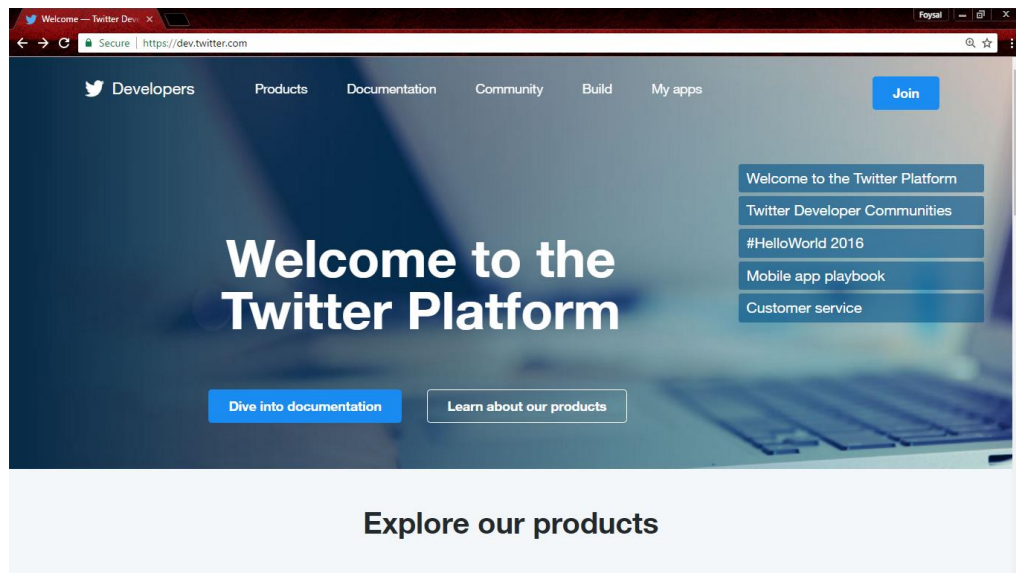
This project will be done using Python and R mainly. The first step of the project is to gather the data that will be required to complete the project. To gather live tweets during the matches, a Python script was created which will run for the full duration of the match to collect every tweet for each game and save to a csv file. A filter function was used (`twitterStream.filter`) so that the Python script can request the twitter database for the tweets with the keywords related to the project. To make it easier for the twitter Streaming API to work python, I have used a Python package called Tweepy. Which made it easier to connect to the twitter database and collect tweets in real time.

For the Python script work and to connect to the Twitter database successfully, we need twitter API keys. The API keys can be obtained by making a developer account on <https://dev.twitter.com/> or if you have a twitter account then your login with your twitter credentials. The Python that was created for the data mining phase can be found in *Appendix B*

```
# Twitter Authentication Info:
ckey = 'msjng3EcYPGdL19gt2npzMnZn'
csecret = 'Kx4DyXylBXnGkdHOJ2Szm2KDyWYSA6ohVm7MgY2qw49wH7sFt'
atoken = '313278116-BOQBIDPukATJAjxbWP4BHGGgXAaf3huylicaxgCP'
asecret = 'eoBSIy91NUVtDLQu0djYZvaPzt9ZiKoKNlfDBFIxMu758'
```

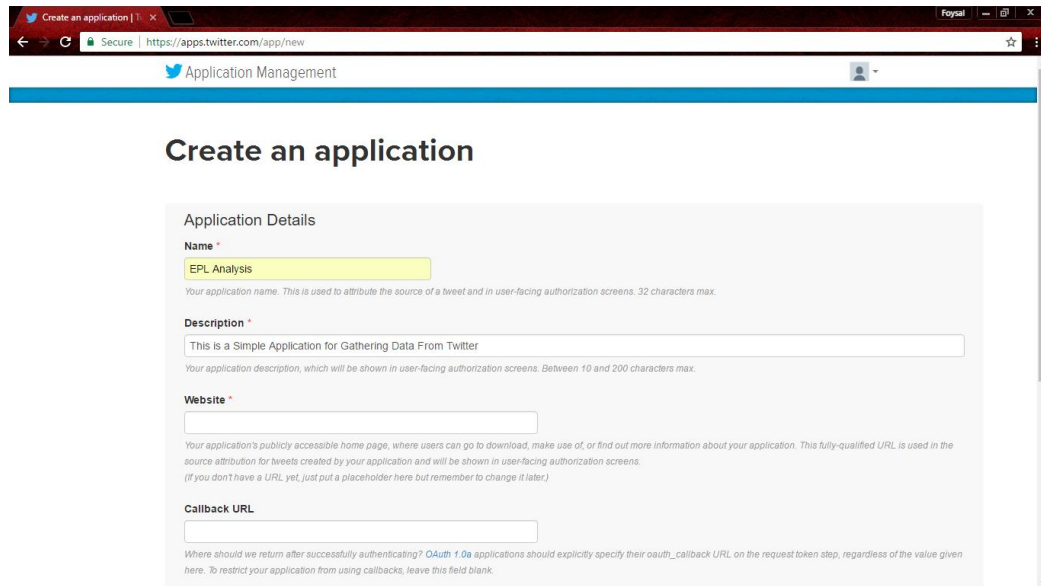
Example of Twitter API keys

The first step of obtaining the API key is to visit <https://dev.twitter.com/> and login using twitter credential or make a new account. This must be done for obtaining API Key.



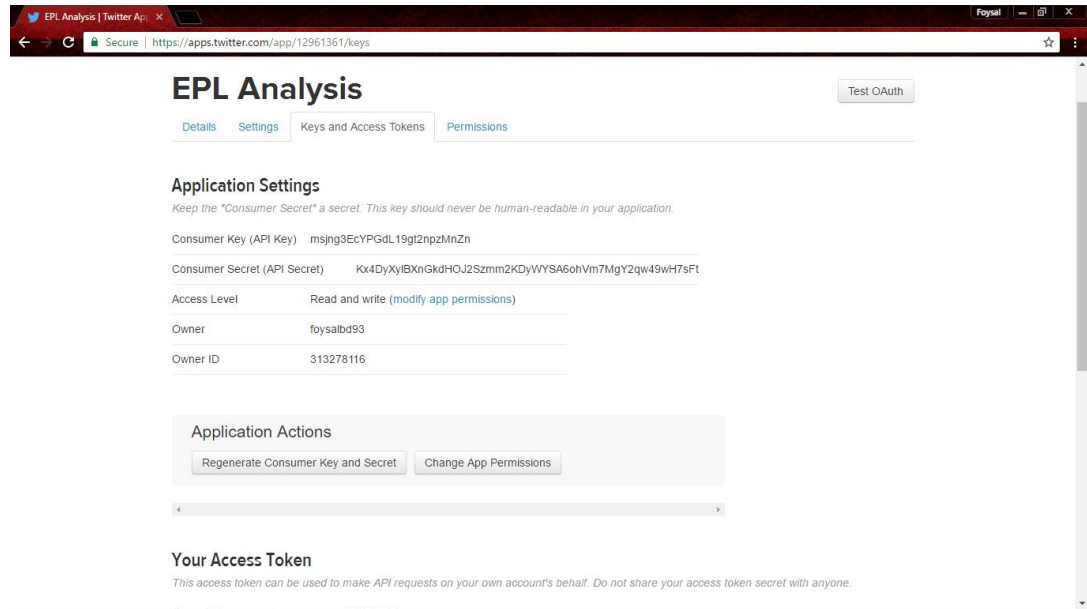
Twitter developer Page: <https://dev.twitter.com/>

Next step is to create an app; this can be done through by clicking My apps at the top of the page. Where the user will be asked to create a new app, then the user will be asked to fill out the application details. Once all the details are filled out, the user can create the app.

A screenshot of a web browser showing the 'Create an application' page on the Twitter developer portal. The browser's address bar shows 'https://apps.twitter.com/app/new'. The page has a blue header with the Twitter logo and 'Application Management'. The main heading is 'Create an application'. Below it is a form titled 'Application Details' with four sections: 'Name' (with a yellow highlight and the text 'EPL Analysis'), 'Description' (with the text 'This is a Simple Application for Gathering Data From Twitter'), 'Website', and 'Callback URL'. Each section has a small text box and a descriptive note below it. The 'Name' note says 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.' The 'Description' note says 'Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.' The 'Website' note says '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.)' The 'Callback URL' note says 'Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.'

Creating a New Application

Once the application is created, the user can see their API key and API Secret, this Key is unique as every user has a different API Key. To make an authorised calls to the Twitter API, API key must be obtained. Once the API Key Obtained, it can be used to gather data from twitter.



Application with the API Key

2.5.1.1 Twitter API:

To get this project up and running, I needed to get data. There are many ways to get twitter get, some data can be historical and some can be current. The efficient way to get the twitter data is through the usage of an API (Application Program Interface), more specifically Twitter API. Twitter has few API, that can be used to get data that required for anyone. The Search API, Streaming APIs and Webhook APIs are some of the examples APIs that twitter offer to its user. For this project, I have used the streaming API to get tweets in real time during the live matches.

2.5.1.2 Data:

The collected data are related to the premier league matches and was collected from 2016/2017 season during the live matches. All the datasets are not the same size as some contains more data than others. Some dataset could be large because of the two teams involved in the match could be rivals and might be a very competitive match. The dataset might be small because of the opposite reasons; it can be less competitive match that not many people watched.

2.5.1.3 Dataset:

There is multiple dataset that was collected during live matches, one team might feature in multiple dataset. Data was gathered by using the twitter streaming API. A python script was created to gather data during the matches and saved a csv file. Twitter streaming API was used in the python script to gather live tweets during the matches. Each premier league matches have its own unique hashtags (e.g. Manchester United Vs Chelsea is #MUNCHE), which is made from a combination of the two team names. I have also used the twitter handle of the two teams involved in the match (@ManUtd and @ChelseaFC). Using the hashtag will allow the API to get tweets that are only related to two teams involved in the match and wouldn't get any other matches tweets. The table below contains the hashtags that were used for to get the tweets related to the matches. The matches highlighted are ones that have been used for the project and the datasets are picked randomly.

Match	Official Hashtags	Date	Score
Manchester City Vs Liverpool	#MCILIV	19/03/2017	1-1
Crystal Palace Vs Watford	#CRYWAT	18/03/2017	1-0
Everton Vs Hull City	#EVEHUL	18/03/2017	4-0
Liverpool Vs Burnley	#LIVBUR	12/03/2017	2-1
Middlesbrough Vs Manchester United	#MIDMUN	19/03/2017	1-3
Stoke City Vs Chelsea	#STKCHE	18/03/2017	1-2
Sunderland Vs Burnley	#SUNBUR	18/03/2017	0-0
Tottenham Hotspur Vs Southampton	#TOTSOU	19/03/2017	2-1
West Bromwich Albion Vs Arsenal	#WBAARS	18/03/2017	3-1
West Ham United Vs Leicester City	#WHULEI	18/03/2017	2-3
Burnley Vs Tottenham Hotspur	#BURTOT	01/04/2017	0-2
Chelsea Vs Crystal Palace	#CHECRY	01/04/2017	1-2
Hull City Vs West Ham United	#HULWHU	01/04/2017	2-1
Leicester City Vs Stoke City	#LEISTK	01/04/2017	2-0
Liverpool Vs Everton	#LIVEVE	01/04/2017	3-1
Manchester United Vs West Bromwich Albion	#MUNWBA	01/04/2017	0-0
Watford Vs Sunderland	#WATSUN	01/04/2017	1-0
Southampton Vs A.F.C. Bournemouth	#SOUBOU	01/04/2017	0-0
Arsenal Vs Manchester City	#ARSMCI	02/04/2017	2-2
Swansea City Vs Middlesbrough	#SWAMID	02/04/2017	0-0
A.F.C. Bournemouth Vs Swansea City	#BOUSWA	18/04/2017	2-0

2.5.1.4 Data Cleaning:

The data cleaning involves cleaning the dataset so that data classification can easier for the machine learning algorithms, this can be done manually by using Microsoft excel or in R studios. Unwanted data is any information that cannot be used by the

machine learning algorithms, this can include punctuation, html links, numbers and words containing any special characters e.g.#, @ etc. Most of the unwanted attributes were removed manually, to perform sentiment analysis only the text or tweets attribute is required. Second part of the cleaning was done in R studio using the `gsub` function. The duplicate data was also deleted by using the `unique` function

2.5.2 Data Analysis:

Analysis 1: Sentimental analysis:

Setting the environment:

Before starting to analyse any dataset, we need to set our working directory, this is where our dataset and output will be saved. We also add the packages that are required, this includes *sentiment* package for scoring tweets based on emotion and polarity, *plyr* package for splitting tweets into sentences for classification and *tm* package which will be used for text mining. Then the lexicons positive and negative word dictionaries are imported into R, the dictionaries will be used by machine learning algorithms to score the data.

```
1  #4TH YEAR DATA ANALYTICS PROJECT--EPL ANALYSIS
2  #FASIAL BASHAR - X13358851-NATIONAL COLLEGE OF IRELAND
3
4  #setting the Working Directory
5  setwd("G:\\Data")
6
7  #loading packages
8  library("gsubfn")
9  library("plyr")
10 library("dplyr")
11 library("sentimentr")
12 library("ggplot2")
13 library("wordcloud")
14 library("stringr")
15 library("sentiment")
16 library("Rstem")
17 library("tm")
18 library("RColorBrewer")
19 library("gridExtra")
20
21
34 #Loading the positive-words and negative-words dictionary (will be used for all the Dataset)
35 hu.liu.pos = scan('G:\\Data\\positive-words.txt', what='character', comment.char=';')
36 hu.liu.neg = scan('G:\\Data\\negative-words.txt', what='character', comment.char=';')
37
```

Then the dataset is loaded, that we will be performing the analysis on There are only two attributes in the dataset, Date.Created and Text. Some of the dirty data have been removed already but there are still small unwanted data which must be

removed, so that data can be analysed. The *gsub* function is used to clean the unwanted variable from the dataset. This includes retweet, @people, html links, punctuations, numbers and unneeded white spaces.

```

42 #DATASET:1 : Manchester City vs Liverpool Match (1-1)
43 MCILIV <- read.csv(file = "MCILIV.csv", header = TRUE, sep = ",")
44
45 #Cleaning the Dataset using gsub() function
46 MCILIV$Text = gsub("(RT|via) ((?:\\b\\W*@[\\w+)+)", " ", MCILIV$Text) # First we will remove retweet entities from
47 MCILIV$Text = gsub("@\\w+", " ", MCILIV$Text) # Then remove all "@people"
48 MCILIV$Text = gsub("http\\w+", " ", MCILIV$Text) # removing all the html-links
49 MCILIV$Text = gsub("[[:punct:]]", " ", MCILIV$Text) # removing all the punctuations
50 MCILIV$Text = gsub("[[:digit:]]", " ", MCILIV$Text) # removing any numbers, only text can analysed
51 MCILIV$Text = gsub("[ \\t]{2,}", " ", MCILIV$Text) # removing any unwanted spaces
52 MCILIV$Text = gsub("^\\s+|\\s+$", " ", MCILIV$Text)
53

```

Once the dataset is cleaned, *unique* function is used for removing any duplicate tweets. A Sample of 1000 tweets was selected randomly, saved in a csv file. Performing machine learning algorithm on large dataset requires high computational power, which wasn't available.

```

54 #Removing the Duplicate data/tweets
55 Unique_MCILIV <- unique(MCILIV)
56 write.csv(Unique_MCILIV, file='G:\\Data\\Unique_MCILIV.csv', row.names = F)
57
58 #loading the Unique Dataset
59 Unique_MCILIV <- read.csv("G:\\Data\\Unique_MCILIV.csv")
60
61 # take a random sample of 1000 from a dataset
62 MCILIV.Sample <- Unique_MCILIV[sample(nrow(Unique_MCILIV), 1000),]
63 #write.csv(MCILIV.Sample, file='G:\\Data\\MCILIV_Sample.csv', row.names = F)
64

```

The text attribute is selected from the dataset, as the Date.Created attributes will not be used.

```

69 |
70 # selectig the Text/tweet only from the Dataset
71 MCILIV_Text = MCILIV.Sample$Text
72

```

Here *Score.sentiment* function is being created, *plyr* and *stringr* package is required for this to work. *Plyr* package is used for combining, splitting and applying data, it involves reducing large problem into smaller pieces and operates on the smaller pieces. Once the its completed all the pieces are put back together. This package is used for breaking the tweets into words from sentences so that it's easier for the *score.sentiment* function to perform. *Stringr* package is used for comparing each word from the tweet to the lexicons dictionaries.

```

73 #score.sentiment Algorithm:|
74 #creating a score.sentiment function
75 score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
76 {
77   require(plyr)
78   require(stringr)
79   # we got a vector of sentences. plyr will handle a list
80   # or a vector as an "l" for us
81   # we want a simple array ("a") of scores back, so we use
82   # "l" + "a" + "ply" = "lapply":
83   scores = lapply(sentences, function(sentence, pos.words, neg.words) {
84     # cleaning up the sentences using gsub():
85     sentence = gsub('[:punct:]', '', sentence)
86     sentence = gsub('[:cntrl:]', '', sentence)
87     sentence = gsub('\\d+', '', sentence)
88     # converting the tweets to lower case:
89     sentence = tolower(sentence)
90     #str_split used for to split the sentences into words
91     word.list = str_split(sentence, '\\s+')
92     # sometimes a list() is one level of hierarchy too much
93     words = unlist(word.list)
94     # comparing the words from the tweets with the positive & negative word dictionaries
95     # match() returns the position of the matched words or NA
96     pos.matches = match(words, pos.words)
97     neg.matches = match(words, neg.words)
98     # we are just looking for TRUE/FALSE:
99     pos.matches = !is.na(pos.matches)
100    neg.matches = !is.na(neg.matches)
101    # TRUE or FALSE will be treated as 1 or 0 by sum():
102    score = sum(pos.matches) - sum(neg.matches)
103    return(score)
104  }, pos.words, neg.words, .progress=.progress )
105  scores.df = data.frame(score=scores, text=sentences)
106  return(scores.df)
107 }

```

Then the text is provided to score.sentiment function to produce score for each tweet, this is combined with Hu.Liu's positive and negative dictionaries. Once the progress bar reaches 100%, the analysis is completed and saved. Then the result is visualized by using ggplot.

```

109 #using the score.sentiment to score the tweets
110 MCILIV_scores = score.sentiment(MCILIV_Text, hu.liu.pos, hu.liu.neg, .progress='text')
111 #saving the scored tweets to csv file
112 write.csv(MCILIV_scores, file='G:\\Data\\MCILIV_scores.csv', row.names=F)
113
114
115 scl<- qplot(factor(score), data=MCILIV_scores, geom="bar",
116   fill=factor(score))+xlab("Sentiment_Scores") + ylab("Tweet_Count") + ggtitle("MCLIV Sentiment Scores")
117 grid.arrange(scl, nrow=1)
118
119

```

Here try.error is created for handling *tolower* function, which returns an error when it cannot map special characters, this can include emoticons. If there are no emoticons then, there isn't any error. Using the try.error to convert text to lowercase with *sapply*. Also, removing any NA's that are found in the text.

```

123 # create missing value
124 try.error = function(x)
125 {
126
127   y = NA #Creating a missing value
128   try_error = tryCatch(tolower(x), error=function(e)e) #tryCatch to handle errors
129   if(!inherits(try_error, "error")) #If there are no error
130     y=tolower(x)
131   return(y) # then the result is fine
132 }
133
134 #using the try.error to convert MCILIV_Text to lowercase
135 MCILIV_Text = sapply(MCILIV_Text, try.error)
136
137 # removing all the NAs in Text data if there are any
138 MCILIV_Text = MCILIV_Text[!is.na(MCILIV_Text)]
139 names(MCILIV_Text) = NULL
140

```

classify_emotion function classifies each word in a tweet into specific types of emotions e.g. surprise, joy, sadness, fear, disgust and anger. This function uses the Naïve Bayes algorithm to classify each word in a tweet.

```

142 #classify_emotion using the Naive_Bayes Algorithm
143 class_emo = classify_emotion(MCILIV_Text, algorithm="bayes", prior=1.0)
144 #get the emotions that gives the best result
145 emotion = class_emo[,7]
146 # Replacing all the NA's with "unknown"
147 emotion[is.na(emotion)] = "unknown"
148
149 #saving the Emotion analysis in a CSV file
150 write.csv(class_emo, file='G:\\Data\\MCILIV_Emo.csv', row.names=F)
151

```

classify_polarity function classifies each tweet as positive or negative, this is also part of the Naïve Bayes algorithm. Both *classify_emotion* and *classify_polarity* uses the Sentiment package which allows tweets to be classified

```

152 # classify_polarity using the Naive_Bayes Algorithm
153 class_pol = classify_polarity(MCILIV_Text, algorithm="bayes")
154 # get the polarity that gives best result
155 polarity = class_pol[,3]
156
157 #saving the polarity Analysis in a CSV file
158 write.csv(class_pol, file='G:\\Data\\MCILIV_Pol.csv', row.names=F)
159
160

```

Once *classify_emotion* and *classify_polarity* was completed, the result of both function was created into a data frame which was then saved csv file. The visualization of the results is in the test and evaluation section. Same algorithms were used for analysing all three datasets.

Analysis 2: Top 4 Predictive analysis:

This is the second part of the project, which also analysis data based on tweets from after Game week 35 (30th and 1st of April 2017). The aim of the predictive analysis is to see if the top 5 teams are happy about their position or the game week result. For this analysis, I have collected 1000 recent tweets for Chelsea, Tottenham, Liverpool, Manchester City and Manchester United. As the premier league season is coming to an end, I wanted to see how the fans are feeling based on the result from that game week. This can tell me if how the fans feel about their team making to the top team. We can see by looking at the Premier League table that Chelsea and Tottenham will finish at the first two positions, but I wanted to find out where the rest of the teams will finish at the end of the season based on the fan reaction. Score.sentiment algorithm is used for scoring the tweets and result will be shown through visualization and table. I will connect R directly to twitter API to access data.

Authorising R to Access Twitter:

First need to install the packages that are required for the project, TwitterR Package provides an interface to the Twitter Web API, ROAuth Package allows authentication using the OAuth to twitter Database for R. The API keys are used as part of the authentication process to get the data.

```
10 #Conttecting to twitter API
11 download.file(url='http://curl.haxx.se/ca/cacert.pem', destfile='cacert.pem')
12 reqURL <- 'https://api.twitter.com/oauth/request_token'
13 accessURL <- 'https://api.twitter.com/oauth/access_token'
14 authURL <- 'https://api.twitter.com/oauth/authorize'
15
16 #my API Keys
17 api_key <- "msjng3EcYPGdLl9gt2npzMn2n"
18 api_secret <- "Kx4DyXylBXnGkdHOJ2Szzmm2KDyWYSA6ohVm7MgY2qw49wH7sFt"
19 access_token <- "313278116-BOQBIDPukATJAjxbWP4BHGGgXAaf3huylicaxgCP"
20 access_token_secret <- "eoBSIy91NUVtDLQu0djYZvaPZt9ZiKoKNlfDBFIxMu758"
21
22 #setting twitter authentication details
23 setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)
24
```

Twitter Authorisation Code with API Keys

Searching for Tweets:

Once R has access to the twitter database, searchTwitter function is being used to search twitter for the tweets by using the hashtags and this will only get tweets related

to a given keyword. For this, I'm gathering tweets by each team's name. A sample of 1000 tweets are being gathered, `resultType` is being used for to filter tweets by most recent.

```
25 #Searching for Tweets related to the Keywords.
26 ManUtd.list <- searchTwitter('#MANUTD',num= 1000, resultType = 'recent')
27 ManUtd.df = twListToDF(ManUtd.list)
28 View(ManUtd.df)
29
30 Chelsea.list <- searchTwitter('#Chelsea #premierleague',num= 1000, resultType = 'recent')
31 Chelsea.df = twListToDF(Chelsea.list)
32 View(Chelsea.df)
33
34 Liverpool.list <- searchTwitter('#LFC',num= 1000, resultType = 'recent')
35 Liverpool.df = twListToDF(Liverpool.list)
36 View(Liverpool.df)
37
38 Tottenham.list <- searchTwitter('#tottenham',num= 1000, resultType = 'recent')
39 Tottenham.df = twListToDF(Tottenham.list)
40 View(Tottenham.df)
41
42 ManCity.list <- searchTwitter('#mancity',num= 1000, resultType = 'recent')
43 ManCity.df = twListToDF(ManCity.list)
44 View(ManCity.df)
45
```

Creating a function to score the tweet, this function is used latter stage for scoring the tweets using the positive and negative word dictionary.

```
46 #Generating the function
47 score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
48 {
49   #requires plyr and stringr package
50   require(plyr)
51   require(stringr)
52   # we got a vector of sentences. plyr will handle a list
53   # or a vector as an "l" for us
54   # we want a simple array ("a") of scores back, so we use
55   # "l" + "a" + "ply" = "laply":
56   scores = laply(sentences, function(sentence, pos.words, neg.words) {
57     # cleaning up the sentences using gsub():
58     sentence = gsub('[:punct:]', '', sentence)
59     sentence = gsub('[:cntrl:]', '', sentence)
60     sentence = gsub('\\d+', '', sentence)
61     # converting the tweets to lower case:
62     sentence = tolower(sentence)
63     #str_split used for to split the sentences into words
64     word.list = str_split(sentence, '\\s+')
65     # sometimes a list() is one level of hierarchy too much
66     words = unlist(word.list)
67     # comparing the words from the tweets with the positive & negative word dictionaries
68     # match() returns the position of the matched words or NA
69     pos.matches = match(words, pos.words)
70     neg.matches = match(words, neg.words)
71     # we are just looking for TRUE/FALSE:
72     pos.matches = !is.na(pos.matches)
73     neg.matches = !is.na(neg.matches)
74     # TRUE or FALSE will be counted as 1 or 0 by sum() function:
75     score = sum(pos.matches) - sum(neg.matches)
76     return(score)
77   }, pos.words, neg.words, .progress=.progress )
78   scores.df = data.frame(score=scores, text=sentences)
79   return(scores.df)
80 }
81
```

Scoring sentiment for each tweet by using the `score.sentiment` function and the lexicon word dictionary. Each team's tweets were scored separately. For the scoring function to work, *sentiment* package is required.

```

93 #scoring each dataset by using the score.sentiment algorithm
94 ManUtd.scores = score.sentiment(ManUtd.df$text, hu.liu.pos,hu.liu.neg, .progress='text')
95
96 Chelsea.scores = score.sentiment(Chelsea.df$text, hu.liu.pos,hu.liu.neg, .progress='text')
97
98 Liverpool.scores = score.sentiment(Liverpool.df$text,hu.liu.pos,hu.liu.neg, .progress='text')
99
100 Tottenham.scores = score.sentiment(Tottenham.df$text,hu.liu.pos,hu.liu.neg, .progress='text')
101
102 ManCity.scores = score.sentiment(ManCity.df$text, hu.liu.pos, hu.liu.neg, .progress='text')
103

```

Finally, all the score is combined using *rbind* function to create a table to output all teams score together. This table is converted to histogram by using *ggplot* function, so that each team's sentiment score can be compared together.

```

126 #Combing All the Teams.Score together for the Table using rbind() fuction
127 all.scores = rbind(ManUtd.scores, Chelsea.scores, Liverpool.scores, Tottenham.scores, ManCity.scores)
128
129 #creating a table with all the teams and the sentiment score.
130 table(all.scores$score,all.scores$Team)
131
132 #Creating a ggplot based on the score for each team
133 ggplot(data=all.scores) + geom_histogram(mapping=aes(x=score, fill=Team), binwidth=1) +
134   facet_grid(Team~.) + theme_bw() + scale_fill_brewer(palette="Set1") # using "Set1" as colors
135

```

2.6 Testing and Evaluation

Analysis 1: Sentimental Analysis:

Testing:

The system was tested by importing the dataset successfully to R after the dataset was cleaned in excel, as before the dataset wouldn't import due to some unwanted attributes being out of place. Once some of the attributes were removed, data imported into R successfully. Then the machine learning algorithm was tested to see if the tweets are being classified, which can be seen by checking the output file.

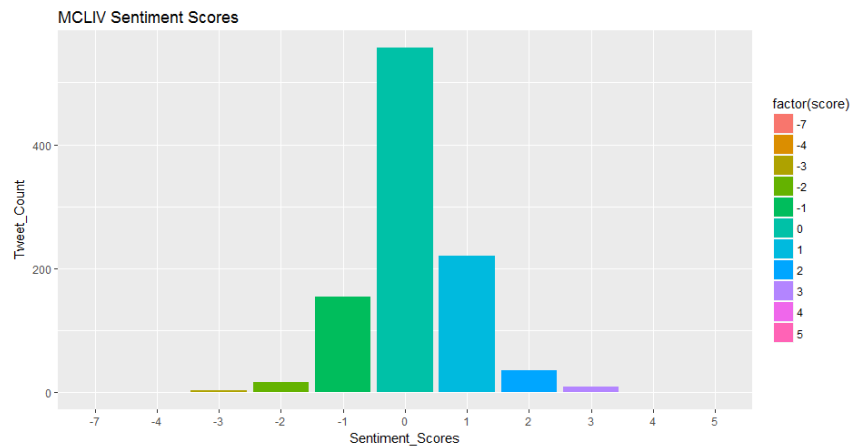
Result:

The tables below display the result of sentimental analysis of the three matches tweets. *score.sentiment* and Naïve Bayes were used for the analysis and the result from both algorithms are shown separately. *emotion_polarity* function result from the Naïve Bayes algorithm is shown in the result, which classifies tweets as “positive” or “negative” or “neutral”. To find the best algorithm out of the two machine algorithms, there is also a table with 100 manual classified tweets. This tells us which algorithm has higher accuracy.

Manchester City vs Liverpool (1-1)

Score.Sentiment:

Tweet Category	Percentage	Number of Tweets
Positive Tweets	26.7%	267
Negative Tweets	17.6%	176
Neutral Tweets	55.7%	557



We can see that score.sentiment scored more than half of the tweets as neutral, positive being the second highest and negative being the lowest.

Naïve Bayes:

Tweet Category	Percentage	Number of Tweets
Positive Tweets	64.4%	644
Negative Tweets	26.5%	265
Neutral Tweets	9.1%	91

The Naïve Bayes algorithm has categorised 64.4% of the tweets as positive, 26.5% negative and 9.1% as neutral.

MCILIV Manual Classification (100 Tweets)

Categories	Manual Classification	%	(NB) Computer Classification	%
Positive Tweets	31	31	60	60
Negative Tweets	21	21	28	28
Neutral Tweets	48	48	12	12

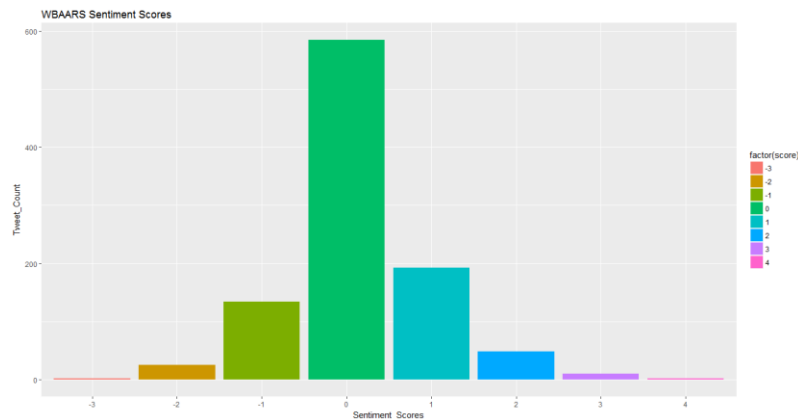
A manual classification was performed to see which algorithm has the highest accuracy, 100 tweets were classified. From the table, we can see that 31% of the tweets were classified as positive, 21% as negative and 48% as natural. From this result, we can see that score.sentiment has similar result to the manual classification, whereas the Naïve Bayes doesn't.

West Bromwich Albion vs Arsenal (3-1)

Score.Sentiment:

Tweet Category	Percentage	Number of Tweets
Positive Tweets	25.4%	254
Negative Tweets	16.1%	161
Neutral Tweets	58.5%	585

score.sentiment shows West Bromwich Albion vs Arsenal tweets have had 58.5% as neutral, 25.4% as positive and 16.1% as negative.



Naïve Bayes:

Tweet Category	Percentage	Number of Tweets
Positive Tweets	69.9%	699
Negative Tweets	21.9%	219
Neutral Tweets	8.2%	82

Naïve Bayes shows that 69.9% of the tweets as positive, 21.9% as negative and 8.2% as neutral.

WBAARS Manual Classification (100 Tweets)

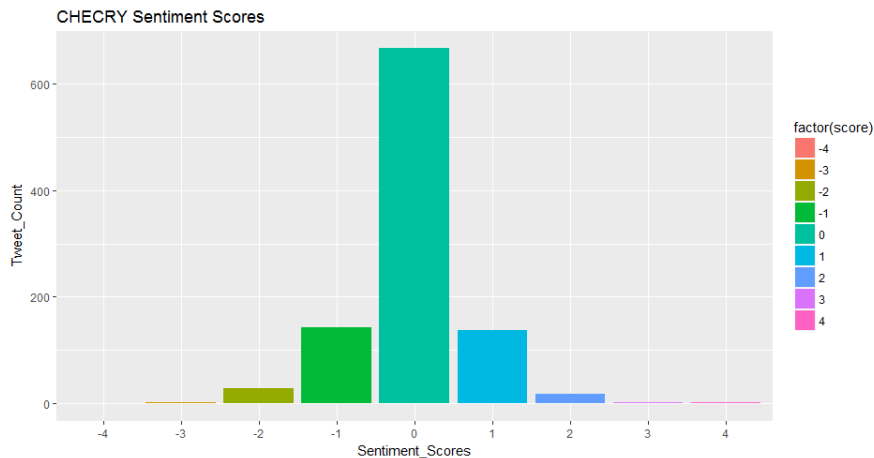
Categories	Manual Classification	%	(NB) Computer Classification	%
Positive Tweets	12	12	68	68
Negative Tweets	33	33	25	25
Neutral Tweets	55	55	7	7

For WBAARS, Manual classification table shows 55% as neutral, 33 as negative and 12% as positive. Again, we can see that this result is similar to the score.sentiment result.

Chelsea vs Crystal Palace (1-2)

Score.Sentiment:

Tweet Category	Percentage	Number of Tweets
Positive Tweets	16%	160
Negative Tweets	17.3%	173
Neutral Tweets	66.7%	667



Score.sentiment shows result contains 66.7% as neutral, 17.3% as negative and 16% as positive.

Naïve Bayes:

Tweet Category	Percentage	Number of Tweets
Positive Tweets	71.5%	715
Negative Tweets	21.7%	217
Neutral Tweets	6.8%	68

Naïve has 71.5% as positive, 21.7% as negative and 6.8% as neutral.

Overall, Based on the result from the score.sentiment function all three match data set has a higher percentage of the tweets as Neutral. The Naïve Bayes algorithm's result contains a higher percentage of the tweets as positive. Manual classification result backs up the score.sentiment result with a higher percentage of the tweets as Neutral. Therefore, score.sentiment outputted the best result.

Evaluation:

In the above tables can see the result of both score.sentiment and Naïve Bayes algorithm's sentimental analysis of the three datasets. Both algorithms were performed to find out how fans were reacting during the match, but the results from the analysis doesn't give conclusive evidence. The two algorithms show different categories of high result, 55.7% - 66.7% of the tweets were categorised as neutral by score.sentiment and 64.4% - 71.5% as positive by Naïve Bayes classification. To get conclusive result, 100 tweets each from two data sets was manually classified to see which algorithm has highest accuracy. From the manual classification, we learned that score.sentiment accuracy is above 75% and Naïve Bayes accuracy is 40%.

The lexicons positive and negative dictionary could be the reason for score.sentiment algorithm high accuracy. As the algorithm breaks down each tweet and score each word based on how positive or negative it is.

Analysis 2: Top 4 Prediction:

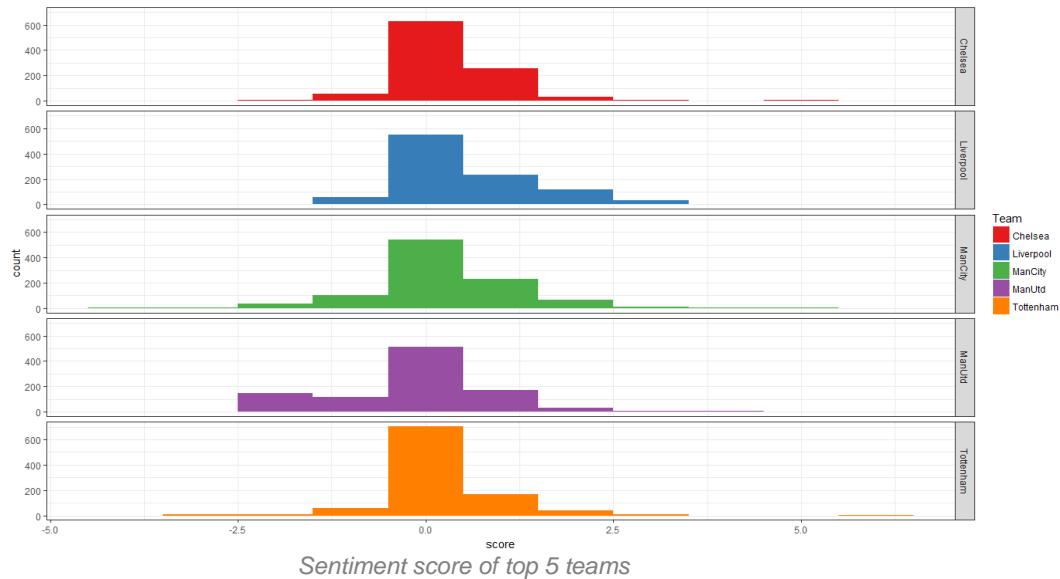
Testing:

The system was tested as the data was gathered using the twitter authentication and API key. Once R was connected to twitter database and begins gathering tweets by keywords, the testing is successful. For further testing the gathered data can loaded into R for analysis. `Score.sentiment` function was tested by scoring the datasets and outputting the result in visualization.

After gathering the 1000 tweets, `score.sentiment` function was created to score the collected tweets, which uses sentiment package. The lexicon word dictionary was also loaded so that it can be used for scoring the tweets. `rbind` function was used to combine all the scores into one table, based on the table Histogram was created.

Result:

Based on the diagram below we can see out of the top 5 sides Chelsea's fans are the happiest, as they beat Everton (0-3). In second it was Tottenham, who has beaten their rivals Arsenal in North London Derby (2-0). Liverpool were third After their narrow away win against (0-1). Followed by Manchester City in fourth with an away draw against Middlesbrough (2-2) and finally followed by Manchester United at fifth with a home draw against Swansea City (1-1). Based one week's data of fans reaction we can see that Chelsea fans are reacting positively to their teams while Manchester United fans are the angriest with high number of negative tweets. When the fans aren't happy with their team performance, some players might not play well due to the negative atmosphere during the match. Based on sample data of 1000 suggest that Manchester United will finish outside the top four, even though they have one more game in hand. As for the top of the table there might be a change in position of teams in third and fourth. Whereas Chelsea doesn't look to be slipping in their race for the premier league title, While Tottenham will finish in second position.



Evaluation:

Based on the result obtained from the sample 1000 tweets, this doesn't tell us how all the fans feel about their teams during the full season just only how they feel about one match, but this does tell us how some fans are feeling towards their team's performance and position on the point table. One factor that is also affecting the result is that the stage of the season. There are only three game weeks left, so some of the team positions might be not affected on the table and some might be. To analyse how the fans' reaction can have influence in their team's performance, larger dataset would be required to get correct prediction. To get the accurate prediction of the points table, not only tweets needs to be analysed but also each team's performance stat from previous seasons. Which would also require a better machine learning algorithm to process larger dataset.

3 Conclusions

When we first looked at the result of both algorithms Naïve Bayes algorithm had a higher percentage of positive tweets compared to score.sentiment which had higher neutral tweets. Once we manually classified data from two out of the three datasets, we noticed that score.sentiment algorithm scored significantly higher accuracy than the Naïve Bayes algorithm. Therefore, at least for this project score.sentiment is much better machine learning algorithm than the Naïve Bayes Algorithm. score.sentiment uses an algorithm that assigns score by counting how many words in a tweet are “positive” or “negative”. Therefore, our result tells us that a higher percentage of the fan reaction was Neutral.

There are other Machine Learning algorithms that can be used for the sentiment and predictive analysis, likes of Decision tree, Random Forest and Support Vector Machine.

4 Further development or research

Given more computational power and time this project can be extended to longer script and more dataset can help to run more analysis. Platforms like Python and Weka can be utilized for better result. Also with better Python knowledge, data mining can be better and higher quality. This can be extended to comparing multiple football leagues dataset and run a more predictive analysis based on football statistics. This can be also evolved into a website where users can see the live result of the analyses that will be carried out, with the help of twitter API. The system could also have more learning algorithm, to improve the system over time. K-means cluster can be used for categorising tweets based on the keywords in a sentence.

The top 4 prediction can also be way better with better knowledge of complex machine learning algorithms. Whereas we are using only twitter data for prediction, but in the future, more accurate prediction can be made with teams past stats used in the analysis.

5 References

- Sentiment Analysis | Lexalytics.* 2016. *Sentiment Analysis | Lexalytics.* [ONLINE] Available at: <https://www.lexalytics.com/technology/sentiment>. [Accessed 10 December 2016].
- R: What is R? 2016.* *R: What is R?* [ONLINE] Available at: <https://www.rproject.org/about.html>. [Accessed 10 December 2016].
- Analytics Vidhya.* 2016. *Perfect way to build a Predictive Model in less than 10 minutes.* [ONLINE] Available at: <https://www.analyticsvidhya.com/blog/2015/09/perfect-build-predictive-model-10-minutes/>. [Accessed 10 December 2016].
- R-bloggers.* 2016. *Sentiment analysis with machine learning in R | R-bloggers.* [ONLINE] Available at: <https://www.r-bloggers.com/sentiment-analysis-with-machine-learning-in-r/>. [Accessed 10 December 2016].
- Predictive Analytics | R-bloggers.* 2016. *Predictive Analytics | R-bloggers.* [ONLINE] Available at: <https://www.r-bloggers.com/tag/predictive-analytics/>. [Accessed 10 December 2016].
- R-bloggers.* 2016. *Twitter sentiment analysis with R | R-bloggers.* [ONLINE] Available at: <https://www.r-bloggers.com/twitter-sentiment-analysis-with-r/>. [Accessed 10 December 2016].
- sentiment - Mining Twitter with R.* 2016. *sentiment - Mining Twitter with R.* [ONLINE] Available at: <https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment>. [Accessed 11 December 2016].
- Anon, (2017).* [online] Available at: <https://dev.twitter.com/rest/public/search> [Accessed 27 Apr. 2017].
- Premierleague.com.* (2017). *Premier League Clubs – Fixtures, Results, Stats & Profiles.* [online] Available at: <https://www.premierleague.com/clubs> [Accessed 27 Apr. 2017].
- R), W., R), W., & Srivastava, T. (2014).* 2014 FIFA WC Winner Predicted Using Twitter Feed (In R). *Analytics Vidhya.* Retrieved 27 Apr. 2017, from <https://www.analyticsvidhya.com/blog/2014/07/world-cheering-2014-fifa-wc-winner-twitter/>
- Docs.tweepy.org.* (2017). *Streaming with Tweepy — tweepy 3.5.0 documentation.* [online] Available at: http://docs.tweepy.org/en/v3.5.0/streaming_how_to.html [Accessed 27 Apr. 2017].
- DBD, U. (2017).* *KDD Process/Overview.* [online] www2.cs.uregina.ca. Available at: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html [Accessed 28 Apr. 2017].
- Leondes, C. (2000).* *Knowledge-Based Systems.* 1st ed. Burlington: Elsevier.
- Papadopoulos, H., Maglogiannis, I. and Iliadis, L. (2012).* *Artificial intelligence applications and innovations.* 1st ed. Heidelberg: Springer.
- Rhandbook.wordpress.com.* (2017). *sentiment analysis using R | R Handbook.* [online] Available at: <https://rhandbook.wordpress.com/tag/sentiment-analysis-using-r/> [Accessed 29 Apr. 2017].
- Liu, B. (2017).* *Opinion Mining, Sentiment Analysis, Opinion Extraction.* [online] [Cs.uic.edu](http://www.cs.uic.edu). Available at: <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html> [Accessed 30 Apr. 2017].
- En.wikipedia.org.* (2017). *Naive Bayes classifier.* [online] Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed 30 Apr. 2017].

Breen, J., (2011). slides from my R tutorial on Twitter text mining #rstats. [online] Available at: <https://jeffreymbreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> [Accessed 30 Apr. 2017].

Miner, G. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. 1st ed. Amsterdam: Academic Press.ss

Paeng Angnakoon. (2013). Mining Twitter with R - Tutorial 2: Scoring tweets. [Online Video]. 12 September 2013. Available from: <https://www.youtube.com/watch?v=S1y3PxULNaQ&feature=youtu.be.com/watch?v=S1y3PxULNaQ&feature=youtu.be>. [Accessed: 1 May 2017].

Paeng Angnakoon. (2013). Mining Twitter with R - Tutorial 3: Scoring tweets. [Online Video]. 12 September 2013. Available from: <https://www.youtube.com/watch?v=S1y3PxULNaQ&feature=youtu.be.com/watch?v=S1y3PxULNaQ&feature=youtu.be>. [Accessed: 1 May 2017].

Paeng Angnakoon. (2013). Mining Twitter with R - Tutorial 4: Scoring tweets. [Online Video]. 12 September 2013. Available from: <https://www.youtube.com/watch?v=S1y3PxULNaQ&feature=youtu.be.com/watch?v=S1y3PxULNaQ&feature=youtu.be>. [Accessed: 1 May 2017].

How to use the Twitter API v1.1 with Python to stream tweets. (2017). YouTube. Retrieved 2 May 2017, from <https://www.youtube.com/watch?v=pUUxmrvl2FE&t=679s>

Python Programming Tutorials. (2017). pythonprogramming.net. Retrieved 2 May 2017, from <https://pythonprogramming.net/use-twitter-api-v1-1-python-stream-tweets/>

R: What is R? (2017). [R-project.org](http://r-project.org). Retrieved 2 May 2017, from <https://www.r-project.org/about.html>

Reddy, V. (2016). *Sentiment Analysis Using R Language | Evoke Technologies*. Evoke Technologies Blog. Retrieved 3 May 2017, from <http://www.evoketechnologies.com/blog/sentiment-analysis-r-language/>

CRAN - Package gsubfn. (2017). [Cran.r-project.org](http://cran.r-project.org). Retrieved 5 May 2017, from <https://cran.r-project.org/web/packages/gsubfn/index.html>

Wickham, H. (2017). *Tools for Splitting, Applying and Combining Data [R package plyr version 1.8.4]*. [Cran.r-project.org](http://cran.r-project.org). Retrieved 5 May 2017, from <https://cran.r-project.org/web/packages/plyr/index.html>

Tse, R. (2013). *tolower() – error catching unmappable characters*. R-bloggers. Retrieved 5 May 2017, from <https://www.r-bloggers.com/tolower-error-catching-unmappable-characters/>

6 Appendix: A

6.1 Project Proposal

Objectives

The objective of the project is to do a sentimental analysis on data that I'll be gathering from twitter during the premier league match at the weekend and get statistical data from external source. The use of external data is to answer some question related to football players and their playing form. The sentimental analysis result will be compared with the statistical result in the graph.

Some questions I'll be trying attempt at the project (Not Limited to):

Which teams have gotten more mentioned on twitter, rank each team.

Which team gets positive or negative tweets and rank them from good to bad.

Compare positive and negative result by visualizing.

Attempt to see if the football fans' tweets have any effect if certain player gets dropped from the team and compare it his stats.

Background

Football is the best sports and most followed sports in the world. Wherever people are in the world, they know about and they follow football. Everyone has that favourite team that they support in their country or some other team in different countries.

The main reason for choosing football to analyse is that I wanted to how people react to a certain team or match. Although there are many top football leagues are around the world, I decided to focus on the English Premier League (EPL). Fans everywhere want their opinions to be heard from other fans and everyone has certain views/opinions on certain matches or certain teams. Social media's like Facebook and Twitter are used by most fans to show their disappointment or excitement after a match of their favourite team that won or lost. Twitter doesn't categorise the tweets as positive or negative, I so thought maybe I can show it though my project.

The aim of this project is to showcase the sentiment analysis with the help of visualization and be able compare if the fans' opinions have any effect on certain player being dropped from the team or is it due to their poor playing performance. This will cross examination of the data that I gather from twitter and player's statistics from online resources.

Sentiment analysis is also known as opinion mining, identifying and categorizing the text is the main objective. The data can be expressed in text form, to determine the user's attitude towards certain topics. Sentimental analysis has become very popular in the marketing area, where a certain organization wants to know positive and negative things people say towards them. There is some powerful data mining software that is available to data scientist.

Technical Approach

- Research further to know how to get data from twitter live by using the twitter API.
- Start to collect the tweets/data and save them on database system, e.g. MySQL, or excel.
- Get the statistical data from external websites (if Available).
- Analyse the data via R or other tools.
- Prepare the data result to be shown in form of visualisation.

Resources required

Python and R Studio

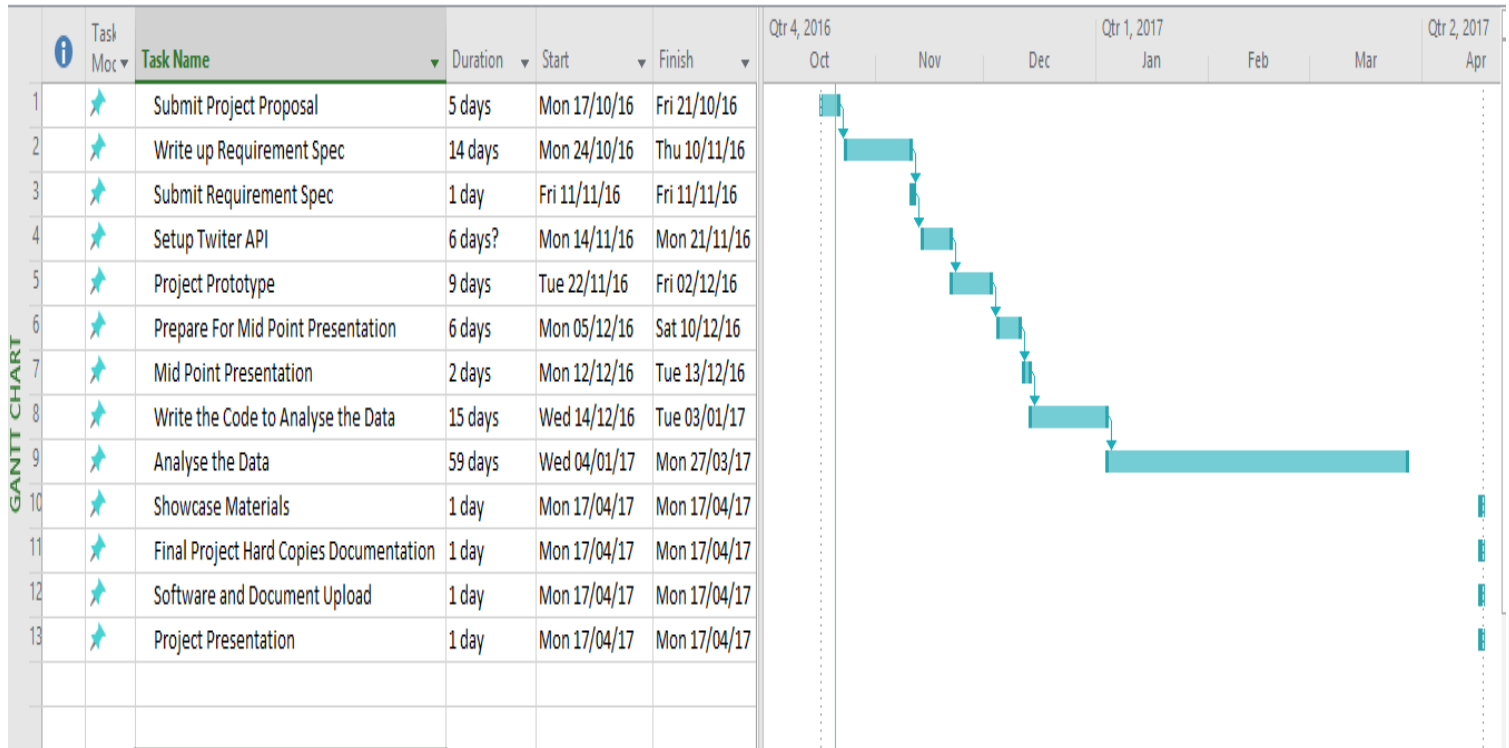
Tweepy package for Python

R Packages – twitterR, Sentiment

Developer Account from <https://dev.twitter.com/>

Project Plan

TASK	PRIORITY	DAYS	START DATE	FINISH DATE	DUE DATE
Project Proposal	Medium	7	14/10/2016	21/10/2016	21/10/2016
Requirement Spec	Medium	10	28/10/2016	8/11/2016	11/11/2016
Get Data	Medium	Throughout project			
Project Prototype	High		05/11/2016	31/11/2016	02/12/2016
Mid-Point Presentation	Medium	1	17/12/2016	17/12/2016	16/12/2016 17/12/2016
Analyse the data	Low	Throughout project			
Showcase Materials	Low	20	20 th March 2017	09 th April 2017	10 th April 2017
Final Project Hard Copies Documentation	Low			10 th May 2017	10 th May 2017
Software & Doc Upload	Low			14 th May 2017	14 th May 2017
Project Presentation	Low			16 th May 2017	16 th May 2017



Technical Details

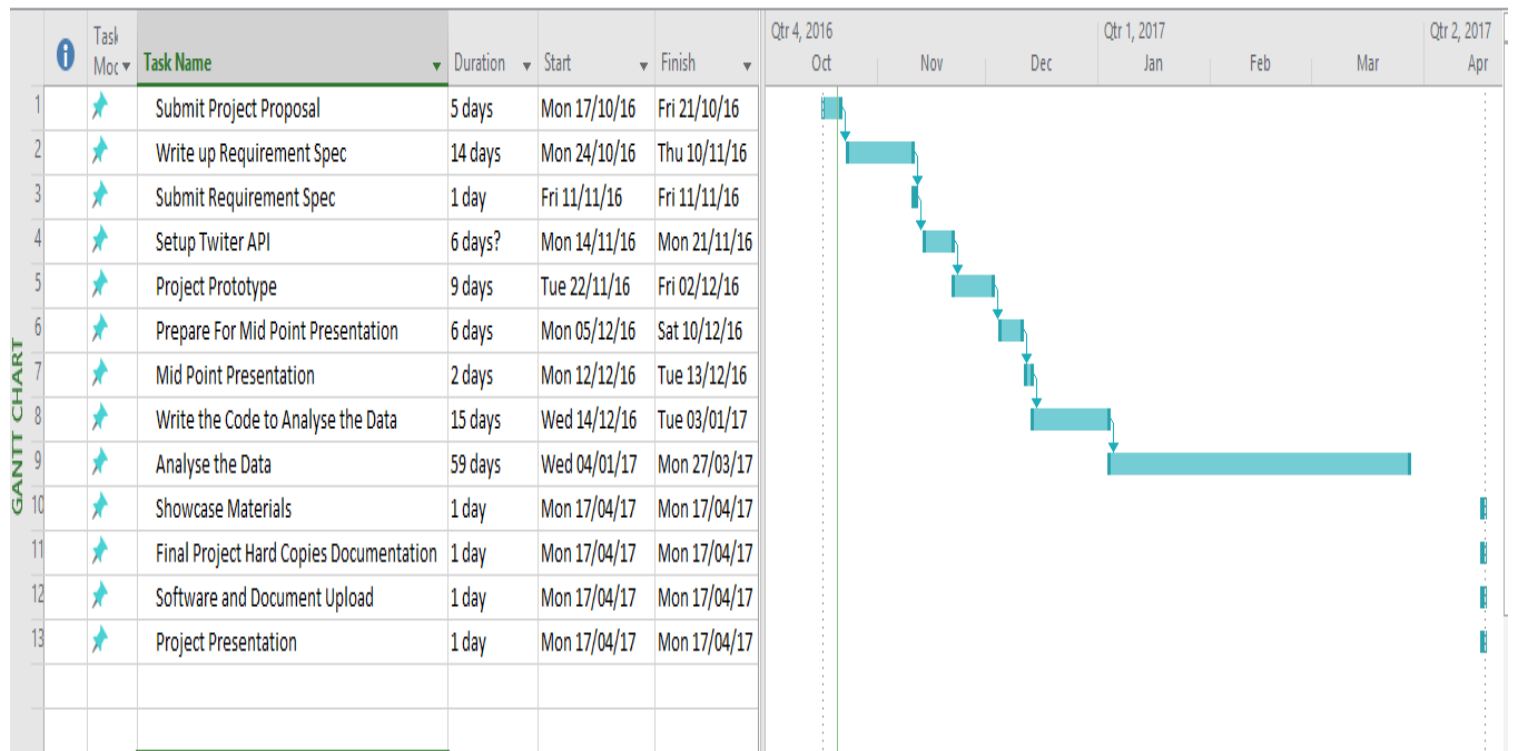
R will be used to evaluate the outcome of the project based on the data that I'll gather data from twitter and get past games statistics from external source. I will be using a database system to save the data that I will mine from twitter through the Twitter API. I will also some external dataset which I'll gather from websites like <http://opisthokonta.net/> , <http://www.football-data.co.uk/> etc.

Evaluation

The system evaluation will be done throughout the project, as I will be gathering the data from Twitter via twitter API. After analysing that data I'll be able to find out if the result meets the project requirement. The test will be carried out as I progress throughout the year, the data's will show with the help of visualization.

Signature of student and date

6.2 Project Plan



6.3 Monthly Journals

Reflective Journal

Student name: Fasial Bashar **Programme:** BSc in Computing (Data Analytics) **Month:** 1 (September)

My Achievements

This month I had to think of project idea related to my chosen stream, which is Data Analytics. I wasn't too sure about what kind of project is related to Data Analytics, as there aren't many examples that I could find online and there wasn't any past project either to look for some guidance. The closest I found a project related to Data Analytics was a project from Business stream. So, I decided to do extensive research online, but only ideas I was getting was to make a dashboard to show live result from a company or some other source of data. From that I got an idea to make an E-commerce dashboard to show live result of a popular online shop (examples include eBay or online shops, but wasn't sure from which I could the data from.) which would focus on different area of the online shop. After thinking for a while if that's the right project or if the lecturer would think if it was a good idea to make an ecommerce dashboard, I decided to have two project ideas at my disposal. The other idea I came up with was to build an informatics dashboard about airlines and airport combined, where I would have popular destinations, frequent flights, cheapest to expensive ticket price, arrival and departure flights and busiest time of year for flights. For this to work I would have to get live data from different websites, But I was told that those ideas wouldn't work for a project in Data Analytics. I wasn't sure about any other kind of data analytics projects, so I decided to go into the project pitch without any ideas for the project. The three lecturers that were there gave me a few suggestions and told me to have a look at the suggested project list that will be on Moodle.

My Reflection

After I get the suggested project list, I'll have to do research on the ones that I interest in and let the lecturer who proposed the project idea know that I'm interested to do that certain project then hopefully they will allow me to do that project. Later then I'll start my project proposal which will be due later in the month.

Supervisor Meetings

No supervisor meeting was held in this month as it's the first month of the project and we weren't assigned any supervisor yet.

Reflective Journal

Student name: Faisal Bashar **Programme:** BSc in Computing (Data Analytics) **Month:** 2 (October)

My Achievements

After the project pitch, I spend a few days to come up with an idea that relates to data analytics. So, I decided to do an analytic project on football, focusing on the English Premier League. I discussed the project idea with one of the lecturers, he gave me positive feedback and gave me some advice how I can make it better. After that, I completed my project proposal and successfully submitted it on Moodle. The past few weeks I have been doing some research to see how I can approach the project and learn how to set up the Twitter API using Python or R language. I have also started my Project Requirement Specification document, which is due on the 11th of November.

My Reflection

If I want to get Data from Twitter through the API, then I need to learn how to get the data. I can use Python or R to set up the Twitter streaming API, which will allow me to grab real-time data. This is new to me so I have been looking through some Python and R examples online that can help me to setup my Twitter API. I also started to learn Python, so that I can use it in some aspect of the project I need to. My supervisor gave me some notes on R Programming, which can help me get more familiar with it as its new language to me.

Supervisor Meeting

Supervisor Name: Muhammad Iqbal
Date of Meeting: 26/10/2016

Items discussed:

- Initial Project Idea and Project proposal
- How to approach the project step by step
- What modelling I'll use for predictive outcome and sentimental analysis?
- The type of Data, I'll need for the project?

Action Items:

- Do research on Twitter Streaming API
- Find out the Size of Data That I will require to complete the project
- Look over the R Notes
- Learn to setup Twitter API with Python or R

Reflective Journal

Student name: Fasial Bashar Programme: BSc in Computing (Data Analytics) Month: 3 (November)

My Achievements

This month, I could plan my project and got some ideas of how I want the end project look but not fully sure yet. I have completed and successfully submitted my project requirement specification before the due date and in the requirement specification document I have set out the functional requirements that I will need to complete for my project, now I have five functional requirements and if I have any additional requirement I will add it to the final document. I also continuing doing research on the tools that I will require to complete the project. I have already started to work on the next document which is technical report that I will need to submit before my mid-point presentation.

My Reflection

For my mid-point presentation, I'm required to make a prototype which I will need to demonstrate for the two lecturers that will be present during the presentation. I'm currently working on the python script that will gather my data for me from twitter, I'm hoping to complete it before the presentation as this will be my prototype. As I only started python introduction in college module, I'm also trying to learn from tutorial and videos from online. Learning python can help me implement the streaming API script for my project, then I can start using R studio and other tools to complete project.

Intended Changes

Next month I'll be able to start coding part of my project and gather the data by using the Streaming API.

Supervisor Meetings

Supervisor Name:

Muhammad Iqbal

Date of Meeting:

26/11/2016 and 09/12/2016

Items discussed:

- Project requirement specification document
- How the requirement specification should be structured?
- Technical report and Mid-Point Presentation.
- What should be in the Report and marking scheme for the presentation

Action Items:

- Complete and submit the requirement specification on the time
- Start the Technical report
- Prepare a porotype for the presentation

Reflective Journal

Student name: Fasial Bashar Programme: BSc in Computing (Data Analytics) Month: 4 (December)

My Achievements

This month, I had my midpoint presentation completed. For the presentation, I developed a work in progress prototype that I had to show to the two lecturers that were present there. The prototype was a python script that will gather live data for me from twitter and save it. After the presentation completed, I was given feedback by the two lecturers and I think the feedbacks will help with me to make my project better. In general, I think the two lecturers were happy my project even though they told me make little changes to the project.

My Reflection

I was happy with the grades that I received for my midpoint presentation, now I just need to keep on working on my project and complete it on time. I haven't done much last few days on the project as I'm currently doing my January exam and have been busy preparing for the exam. As soon as I'm done with the exams I'll continue working on the project.

Supervisor Meetings

This month there were no supervisor meeting, as I had my midpoint presentation. The college was closed for the holiday. My supervisor also gave me feedback during the presentation.

Reflective Journal

Student name: Fasial Bashar **Programme:** BSc in Computing (Data Analytics) **Month:** 5 (January)

My Achievement

Everything is going fine, had a little bit of problem with getting the data that I'm required for the project. The problem was that, when I try to collect data from twitter I also get unwanted data that is not required for the project. I discussed that problem with my supervisor and he gave some advice on how to fix that problem. Currently working on the next stage of the project. I also I met with my supervisor for mid-point presentation feedback.

Supervisor Meeting

Supervisor Name: Muhammad Iqbal

Date of Meeting: 19/02/2017

Items discussed:

- General feedback of mid-point presentation
- Progress of the project
- The problems that I'm having with my code

Action Items:

- Work on the technical aspect of the project
- Add more calculations

Reflective Journal

Student name: Faisal Bashir Programme: BSc in Computing (Data Analytics) Month: 6 (February)

My Achievements

Collected all the data that I'm required for the project, using Tweepy library in Python. While collecting the data, I also gathered some unwanted data. Some of the popular team match data files are bigger than some of the smaller teams. Difficult part of this will be cleaning the data, as there are too much dirty data. I cleaned some parts of the data (Attributes) manually and the rest of the cleaning will be done through R Studio. Once the data was cleaned, the file got smaller. I have collected data of multiple matches, but I will only use the few of the dataset. so, I can compare how different algorithm scores each tweet and compare to the other matches.

Next part of the project will be the important part, as I will be attempting to use the algorithms on the tweet. The two algorithms that I'm hoping to use for the sentimental analysis are Naïve Bayes and score.sentiment. Then I start working on the final report.

Reflective Journal

Student name: Faisal Bashar Programme: BSc in Computing (Data Analytics) Month: 7 (March)

My Achievements

Cleaning some parts of the dataset by using Microsoft Excel, I started using R studio for the next phase of the project. I have loaded the datasets I need for the project, I started next phase of cleaning the dataset this includes removing hashtags, html links and numbers as the sentimental analysis only works on text. Cleaning the dataset is an important part of the project as dirty data can have effect on how the tweets are scored in sentimental analysis. I have started running sentimental analysis on the datasets. I'm using 1000 random tweets from each of the selected data set to perform my analysis. Too many tweets take longer and more computational power to analysis, so I have decided to use the *sample* function in R to select 1000 tweets randomly.

My Reflection

I felt, that cleaning the dataset using excel was good, but if I wanted the data to be ready for the analysis, then I would need to clean it in R using the *gsub* function that will remove the unwanted data.

Intended Changes

Once my analysis is done, I can start writing up the final report. Which will show what kind of algorithms were used to perform the analysis. This will also include findings and conclusion for each dataset.

6.4 Appendix B: Python Script for Data Mining

```
#OnlineResource: https://www.youtube.com/watch?v=pUUxmrv12FE&t=679s
#https://pythonprogramming.net/use-twitter-api-v1-1-python-stream-tweets/

#importing packages:
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener, json
from pprint import pprint
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener

# Twitter Authentication Info:
ckey = 'msjng3EcYPGdL19gt2npzMnZn'
csecret = 'Kx4DyXylBXnGkdHOJ2Szmm2KDyWYSA6ohVm7MgY2qw49wH7sFt'
atoken = '313278116-BOQBiDPukATJAjxbWP4BHGGgXAaf3huylicaxgCP'
asecret = 'eoBSIy91NUVtDLQu0djYZvaPZt9ZiKoKNlfDBFIxMu758'

#Creating a Class:
class listener(StreamListener):

    def on_data(self, data):
        line_object = json.dumps(data) # Dumps takes an object and produces a string:
        file = open('MCILIV.csv', 'a') # Creating a .csv file to save data
        file.write(line_object+"\n") # creating new line for each data
        file.close() # close after process complete
        print(data) # print data once connected to TwitterDB
        return(True)

    def on_error(self, status): # Show error if not connected
        print(status)

auth = OAuthHandler(ckey, csecret) #Twitter Authentication
auth.set_access_token(atoken, asecret)

twitterStream = Stream(auth, listener())
stream = Stream(auth, listener())
twitterStream.filter(track=["MCILIV", "ManCity", "LFC"]) #gathering data that have the
given twitter hashatag in it...
```

6.5 Appendix C: Manual Classification

6.5.1 MCILIV

text	Manual Classification	Classification	Correct or Incorrect
second half city get us under way again cityvlf mcfc n n ud d udd ud d udd t co vntklwrt	neutral	negative	incorrect
karakanyasoka wa vokoli maragoli ni game poa wazito natakia ti u	neutral	negative	incorrect
um	neutral	positive	incorrect
we bottled	neutral	positive	incorrect
semoga musim depan klopp eveluasi lagi dan lfc punya striker bertipe n perusak prahara rumah tangga lawan t co sbtimr bcd	neutral	positive	incorrect
living life on the edge defensively	neutral	positive	incorrect
pep guardiola u cthis is one of the happiest days of my life as a manager how we played against liverpool means a l u	positive	positive	correct
liverpool fc manchester city liverpool follow live coverage t co hvigdtgup lfc	neutral	positive	incorrect
come on you red men	positive	positive	correct
was vital not to lose that	neutral	neutral	correct
it has everything but goals so far ud d udd ud d udd t co qx otj	positive	positive	correct
what a bloody tackle from milner fernandinho is comical mcfc mciliv lfc	negative	negative	correct
great game so open	positive	neutral	incorrect
what a goal that would ve been mciliv	positive	positive	correct
le rateeeeeeeeeee de l aneeeeeeeeeeeeeeeeee n mciliv	neutral	positive	incorrect
not a penalty because reasons lfc	neutral	negative	incorrect
james milner scored liverpool s th penalty in all competed recorded at lfc n n t co vphrz cw h	neutral	negative	incorrect
it s been a fine first half	positive	positive	correct
great thing about watching a game at etihad on tv is that you don t understand a word that the adverts say chinese	negative	positive	incorrect
it s boringmilner	neutral	positive	incorrect
vinner begge hengekampene	neutral	positive	incorrect
det er meg helt uforst u e elig hvorfor manchester city supportere piper p u e milner han gjorde en solid jobb for de i u e r mye p u e benk mciliv	neutral	positive	incorrect
mciliv bonus points as it stands n n u toure n u clichy n u coutinho	neutral	negative	incorrect
what the heck is she putting inside n mciliv n t co z q dejgb	negative	negative	correct

best possible way for milner to silence the boo boys lfc lead at the etihad	neutral	neutral	correct
come on city ud d udc t co rrt fz j	positive	positive	correct
ht manchester city liverpool not the best not the worst big second half coming up lfc mciliv cityvlfc t c u	neutral	neutral	correct
bring in bravo we need goals mciliv	positive	negative	incorrect
mciliv mcfc cityvlfc lfc t co e ddd r b	neutral	positive	incorrect
got to be emre can s best performance for lfc thought he was immense today	positive	positive	correct
ndeeep in my heart	positive	positive	correct
lollolna	neutral	positive	incorrect
roses are red	neutral	positive	incorrect
u starfootballleague n u mciliv n u a muhc n u b stephen oday n u c drpaul yahweh video n n wat trndnl t co zx a ks	neutral	negative	incorrect
mane with some insane pace	positive	negative	incorrect
fucking hell i m sitting in the gym and mciliv is on and the people here just went wild when lallana bottled it to make it ud d ude	negative	negative	correct
a good battle ud d udd ud d udd t co y qkgj gg	positive	neutral	incorrect
city will not finish in the top period you can take that to the bank	neutral	neutral	correct
what the heck is she putting inside n mciliv n t co e qnkfuwqy	negative	negative	correct
milner to liverpool free n nsterling to city u a m n nlfc might lose their young potential player	positive	positive	correct
cabellero doing a dwight gayle here and turning up against lfc	positive	positive	correct
ud d udcf sergio aguero celebrates man city s equaliser against liverpool n nit s with mins remaining mciliv t c u	positive	positive	correct
ft ud d udd ud d udd n nan absorbing game at the etihad as the points are shared cityvlfc mcfc t co s xfbiweta	neutral	negative	incorrect
she s beautiful n mciliv n t co xlhxhtjhgr	positive	positive	correct
u e u e u e u e u e u e a u e u e u e u e c u e d u e u e u e u e u e u e d u e u e u e a u e c u e u e a u e u e u e u e u e a u e d u e u e c u e u e u e u e u e u e u e u e u e u e a u e u e u e u e c u e u e u e b u e u e d u e u e u e u e u e u e u e u e u e u e a u e u e u e u e c u e a u e d u e a u e u e u e d u e u e u e u e u e u d d u d e u d d u d e u d d u d e lfc	neutral	positive	incorrect
sergio aguero cettte apr u e s midi mciliv t co mjynxiipdj	neutral	positive	incorrect
ud d udcf james milner makes it ufe f u e out of ufe f u e from the u aa ufe f this season n nit s man city liverpool mins mciliv u	positive	positive	correct
yessss lofey t co idlxvbdutq	neutral	positive	incorrect

toure should be off	negative	positive	incorrect
d u e couvrez le jardin m u e morial de l etihad stadium u e manc City amp la grande tradition des supporters anglais ud d ude f u bd ufe f ud c uddec ud c udde u	neutral	positive	incorrect
missarnas match fortsatt obese grat mot topp sex i alla fall lfc	neutral	neutral	correct
betvictor new custom nbet u a get n football golf epl sfc n lfc mu fc mcfc afc nentry k golden goal claim u t co ys a wqrwi	neutral	negative	incorrect
u a u u u f u u a u u u u u u f u a u sauditrendat n cityvlf n u u a u u a u u u u u u u a u u u u u u u a u f n u u u u u a u a u u u u u f u u u u a u u u a n mciliv u u a u u u a u u u a u t co tuxookwfst u	neutral	positive	incorrect
can playing well so far looking to add a few extra k lfc	positive	positive	correct
what the heck is she putting inside n mciliv n t co c arpoxy n	negative	negative	correct
what the heck is she putting inside n mciliv n t co iaghlwire	negative	negative	correct
get in there milner lfc ynwa	neutral	negative	incorrect
gooooooal milnerrrr t co rfd b qsuwy	positive	positive	correct
lfc relief for the reds n nsterling can t finish from a cross and neither can fernandinho on the follow up n n	neutral	neutral	correct
he s playing bad	negative	negative	correct
each	neutral	positive	incorrect
feckin c mannnnnnnnnn lfc	negative	positive	incorrect
can is freaking amazing he did a lot of work both in defence and offense lfc	positive	neutral	incorrect
u yellow card man city nman city liverpool n mciliv t co yn hb kg qg	neutral	positive	incorrect
proper game of football that take a point lfc	neutral	neutral	correct
you should be a comedian t co bcbfcvkima	positive	positive	correct
good to see you n n ud d udd ud d udd t co oirveylij	positive	positive	correct
u n n t co is rawtnf	neutral	positive	incorrect
matches between premier league s big slightly irrelevant league table n n lfc n cfc n thfc n mcfc n mu fc u	negative	negative	correct
t co s qsm xpol	neutral	positive	incorrect
rt mcfc vs lfc was a match of missed opportunities	negative	positive	incorrect
man utd have an edge against us ud d ude c t co og b r ou	positive	positive	correct
yes great shot by milner love how klopp doesn t watch penalty shots lfc	positive	neutral	incorrect
non sbaglia il rigore il capitano del lfc come sempre rigore ottenuto da firmino	neutral	positive	incorrect
what the heck is she putting inside n mciliv n t co pr um nab	negative	negative	correct

ntupo ndaaaaaani sana mie na ubavu wangu monicah outer matongo livepool u	neutral	positive	incorrect
firmino dribble and passing has been superb but the finishing was a very poor ud d ude ud d ude mciliv	negative	neutral	incorrect
two penalties not given to liverpool mciliv	neutral	positive	incorrect
fuck off you useless cunts t co fgsjzjxcw	negative	negative	correct
she s beautiful n mciliv n t co l ebkeyxl	positive	positive	correct
she s beautiful n mciliv n t co ekyxyj zgj	positive	positive	correct
ptain dommage sadio mane il aurait d u fb mieux faire mciliv	neutral	negative	incorrect
u e u e u e d u e u e a u e a u e u e u e u e a u e c u e u e u e u e u e u e u e u e b u e u e u e u e u e u e u e u e l f c	neutral	positive	incorrect
ud d udcaa t co cg fxmzxsx	neutral	positive	incorrect
what the heck is she putting inside n mciliv n t co iaghlwire	negative	negative	correct
firmino going off to slide in front of the lfc fans then looks back hahahaha n n t co x jrsunzpd	neutral	positive	incorrect
a solid shift from both sides ud d udd ud d udd t co exgngruq y	positive	positive	correct
pep guardiola u cthis is one of the happiest days of my life as a manager how we played against liverpool means a l u	positive	positive	correct
who turned the penalty feature off then michael oliver having a worldly as usual ud e udd lfc	negative	negative	correct
a solid shift from both sides ud d udd ud d udd t co exgngruq y	positive	positive	correct
klopp	neutral	positive	incorrect
mcfc fans booing james milner mciliv t co vtgfhsmf	negative	positive	incorrect
james milner with a superb penalty on his return to mcfc one of the best players on the pitch today mciliv lfc	positive	positive	correct
goal manchester city liverpool james milner scores a penalty after being booed all game from the city fans m u	neutral	negative	incorrect
how is that not a penalty mciliv	negative	negative	correct
watch manchester city vs liverpool premier league live stream t co ypwa ag pd n n mcfc cityvlfc lfc pl	neutral	positive	incorrect
mciliv grosse intensit u e dans ce choc de pl matip vient mettre sa t u eate sur coup franc u e a passe u e c u f t u e u	neutral	negative	incorrect
never a penalty once you get megged like that	negative	negative	correct
emre and gini were the best for us today milner came next n lfc	positive	positive	correct
win you frauds	negative	positive	incorrect

6.5.2 WBAARS

text	Manual Classification	Classification	Correct or Incorrect
arsenal fans dey laugh us say we dey europa league	neutral	negative	incorrect
arsenal facing th defeat in premier league games nonly games won in the last month both against th tier teams	neutral	negative	incorrect
arsenal fan by any chance xd t co e hg l bxx	neutral	positive	incorrect
awful	negative	negative	correct
arsenal are a mess ud d ude	negative	negative	correct
even bigger twat t co v jnhagcur	negative	positive	incorrect
arsenal should have had a penalty there	negative	negative	correct
ud e udd t co izi uzyoqv	neutral	positive	incorrect
arsenal fans he s won more than you	positive	positive	correct
mins west brom arsenal t co z nyx ojo	neutral	positive	incorrect
ft west brom arsenal now we wait	neutral	positive	incorrect
arsenal ha perdido de sus u faltimos partidos por premier algo in u e dito desde que ars u e ne wenger es el dt	neutral	positive	incorrect
maybe arsenal is just enugu rangers that is across the atlantic	neutral	positive	incorrect
and so the embarrassment continues arsenal	negative	negative	correct
lol arsenal hahahahaha	negative	positive	incorrect
if liverpool win tomorrow they will go points clear of arsenal in th place lfc	neutral	positive	incorrect
goal west brom arsenal n nsub robson kanu stabs the hosts into the lead a minute after coming on u	neutral	positive	incorrect
arsenal have won just two games in the past month against sutton and lincoln	neutral	positive	incorrect
arsenal mah suka gitu ahh	neutral	positive	incorrect
arsenal tv is gonna be scenes ud d ude ud d ude ud d ude ud d ude	neutral	positive	incorrect
wenger out thats it n savearsenal	negative	positive	incorrect
the date is june narsenal finished th behind tottenham nand lost fa cup semi final to city nwenger signs new u	negative	negative	correct
should wenger stay	negative	positive	incorrect
kanu s chance saved by ospina after the striker came one on one wbaars theplshow	neutral	positive	incorrect
lailai our case is different tomorrow n nwe are winners t co rgbcaved q	positive	positive	correct
west bromwich albion vs arsenal saat half time t co ikbft rpi t co yuna xv jr	neutral	positive	incorrect

[illegible]

i kinda	positive	positive	correct
t co pxfea vcta	neutral	positive	incorrect
u a gooooooooooooooooooooool del west bromwich robson kanu	neutral	positive	incorrect
um	neutral	positive	incorrect
the arsenal way	neutral	negative	incorrect
that moment when you re ready to watch arsenal fan tv ud d udc ud c udf f t co zvegrjaijc	neutral	neutral	correct
uff el arsenal otra vez	neutral	positive	incorrect
today s first time it feels like arsenal are back in mid s rudderless and directionless endgame for wenger surel u	negative	positive	incorrect
we will not allow you sack wenger n nwe will not take it he is a philanthropist	positive	positive	correct
ud d udcf back on track and onto points n ncome on you baggies n n wbaars wba t co p dix ac	neutral	positive	incorrect
arsenal ud e udd	neutral	positive	incorrect
ape jadi ngan arsenal nih ozil x dekkk x mainnn	neutral	negative	incorrect
well	positive	positive	correct
any reason why our best players weren t on the pitch in the first place arsenal wengerout ramseyout kronkeout	negative	positive	incorrect
liverpool try their hardest to be the banter club but arsenal are usually like u hold my beer u	neutral	neutral	correct
felicitaciones	neutral	positive	incorrect
arsenal looooooooool	negative	positive	incorrect
arsenal fans planning to do this in the next game t co spahr cxii	negative	negative	correct
arsenal have lost three away league matches in a row for the first time since august october wbaars t co h v u	negative	negative	correct
we will not allow you sack wenger n nwe will not take it he is a philanthropist	positive	positive	correct
in	neutral	positive	incorrect
all for arsenal fan tv	neutral	positive	incorrect
deadddddddd west brom is using arsenal to test run teenagers	positive	positive	correct
totally need to absolutely batter the arsenal twitter account club have to get the message	negative	negative	correct
a manchester united win can move them from th tomorrow	neutral	positive	incorrect
life is too short for me to be stressing about arsenal like this	negative	positive	incorrect
t co xekxrbe ze	neutral	positive	incorrect

ud d ude f t co s waoiwqcm	neutral	positive	incorrect
arsenal soooo poor ud d ude	negative	negative	correct
arsenal fans have been screaming th from th place loool	negative	positive	incorrect
udah gak ada mental lagi ini gua rasa wengerout	neutral	positive	incorrect
on me souffle dans l oreillette que pdnt que je m amuse au salon du fitness	neutral	positive	incorrect
what have we become ud d ude arsenal it s a sad day when we get beat by west brom no one fears arsenal anymore wengerout	negative	negative	correct
then make your voice heard amp demand change t co zi pacg l	negative	positive	incorrect
t co nxmlb vw v	neutral	positive	incorrect
tumben banget update soal bola ud d ude c ud d ude c t co kb e aw y	neutral	positive	incorrect
all because you want united to remain in the th position smh t co ntv mmsmv	neutral	positive	incorrect
u u a u u u a u u u u u u u u a u u u u u u c u u b u	neutral	positive	incorrect
do the honorable thing arsene and	positive	positive	correct
this one wants to occupy arsenal ud d ude d ud d ude d ud d ude e t co xvndww coe	neutral	positive	incorrect
most damning aspect of this is arsenal can t even remotely feel hard done by technically	negative	negative	correct
this is getting plane ridiculous t co uiywngtuo	negative	negative	correct
we really are embarrassing	negative	negative	correct
hahahhahahahah arsenal hahahhahaha	negative	positive	incorrect
arsenal	neutral	positive	incorrect
two craig dawson headers seal an impressive home win for tony pulis s men wbaars t co nojdcgzua	positive	positive	correct
have arsenal stopped saying it s up to arsene to decide when he goes hope so it s not usually up to the employee to decide these things	negative	positive	incorrect
plisss bilang lah wenger saya tdk melatih arsenal lagi di musim depan	neutral	positive	incorrect
that s it arsenal fall to	negative	negative	correct