Abstract- In recent years, Twitter has become a very important and prominent part of people's life. The main reason people use twitter is to share information and their experiences. While expressing their sentiments via tweets they use a certain hashtag for the tweets they publish. When other people agree to it they retweet it, mark it as favorites and even share the tweet. When many people do this for the same hashtag, it starts to trend and the hashtag becomes popular.

In this project, we look onto four hashtags of which three is related to football stars and one related to an award named Ballon d'Or that is given to the best football player of the year. We harvest the data for three footballers and Ballon d'Or award using hashtags. We analyze the data's like tweet text, retweets, favorite counts, the user who publishes those tweets, and hashtags that have been used while publishing tweets. We check the tweets that have been created before the award ceremony and also the tweets that are published after the ceremony. We make 4 buckets each for 4 hashtags and make two categories like:- before award and after awards. To analyze for popularity index we extract only those tweets where the retweets counts and favorites counts are more than 1500. After getting the popular tweets for each hashtag we clean the tweet text and analyze the sentiments of people associated with those tweets. We analyze sentiments for both before awards and after awards scenarios.

Keywords:- twitter, prominent, sentiments, tweets, hashtag, extract, popularity index .

1. Introduction

Twitter has experienced major growth in the last few years. It won't be wrong if we mention it as a revolutionary step in the field of internet and social media [1]. Twitter is becoming a prominent part of people's life as it is used to share information and experiences among its users. According to [2], twitter has 330 million monthly active users 134 million daily active users.

These numbers are increasing rapidly as people are becoming more familiar with it and have become part of their livelihood. In a morning a person might forget to go washroom but not to check the phone or open the twitter. Nowadays, many big companies are investing millions of dollars to do advertising and marketing on twitter [3].

The main reason is that it can reach millions of people around the world and that too in a short period. People also express their feelings and sentiments on twitter via their tweets [4]. Sentiments are just feelings which can be classified in various like:-

1. Fear
2. Joy
3. Love
4. Happiness
5. Anger
6. Disgust
7. Surprise etc.

But here in our project, we are just trying to classify sentiments in two parts:-

1. Positive
2. Negative

1.1 Problem Statement

Here, in this project, we are trying to analyze the twitter data for its popularity index based on retweet count and favorites count. We will only take those tweets of which retweet count and favorites count are more than 1500. Now, these tweets texts are used to analyze the sentiments associated with those tweets.

1.2 Objective

The objectives of this project are:

- To create the developer's app for twitter and connect twitter and python using signature keys to harvest data from twitter by using hashtags like #Ronaldo, #Messi, #Vandijk and #Ballondor(harvest 100 latest tweets).
- To analyze the popularity index of data only those tweets that retweet and favorites count greater than 1500 are taken into consideration.
- Since the tweet text is unstructured, it is first cleaned before processing it further.
- Then we do sentiment analysis on those tweets and know how many positive and negative tweets are attached with certain hashtags before and after the award.

1.3 Existing System

There are various ways that already exist to harvest the data from twitter like using R Studio, JAVA, and even Python. There are also various Machine Learning algorithms to do sentiment analysis on text data. The algorithms like Naïve Bayes' classifier, K- means Clustering, SVM, Neural Network, etc. are present.

1.4 Proposed System

In the proposed system we are using Python and use python library called tweepy and add signature keys in our code to harvest data from twitter.

Those data are stored as .csv file in local machine. We use this data first to analyze the popular tweets by printing those tweets that retweet and favorites count is more than 1500.

Then we use the ML algorithm called SVM to perform sentiment analysis on those tweets. We perform sentiment analysis on the tweets which were created before and after the award ceremony. Then we compare the sentiments of people before and after the award. We try to build a more efficient and powerful model for performing sentiment analysis using SVM.

2. Literature Survey

Literature Survey gives the surface view on the topic and related work that has already been published in some journals or on the internet.
This section focuses on various literature surveys proposed by various people what they got from their personal research.

1. The ascent and improvement of the Internet and the World Wide Web have given a worldwide system to sharing data and working together in confiding seeing someone. No sweat of openness, they have multiplied in our lives to a degree where the clients can get to/share data anyplace whenever [5].
2. Online networking is ruling the universe of showcasing advancement these days. This makes it the ideal stage for any advertising advancement in any field. One of the most

significant parts of online networking promoting is the two route plausibility of correspondence [6].

3. In the past few years, social media sites have experienced tremendous growth. People spread information, opinions, announcements, and behaviors via social media. We have a lot of talking about themselves and their experiences. They also exchange details about how we feel, in particular [7].
4. For the most part, the sack of-words approach has been utilized for mining estimations on the web. In this methodology, singular words are considered rather than complete sentences. Customary AI calculations, for example, Support Vector Machines, Naive Bayes' and Maximum entropy, and so on are ordinarily used to take care of the characterization issues [8].

3. Design & Implementation

This project is divided into five phases and it can be explained as below:-

1. Data Collection
2. Analyzing popularity index
3. Data Cleaning
4. Performing Sentiment analysis
5. Data Visualization

First we show the design of overall project as belows:-

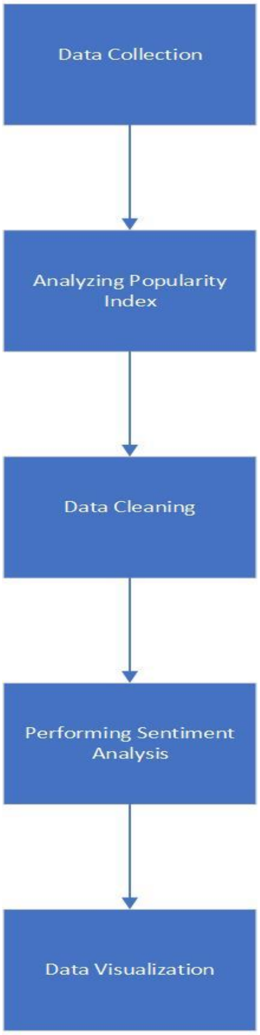
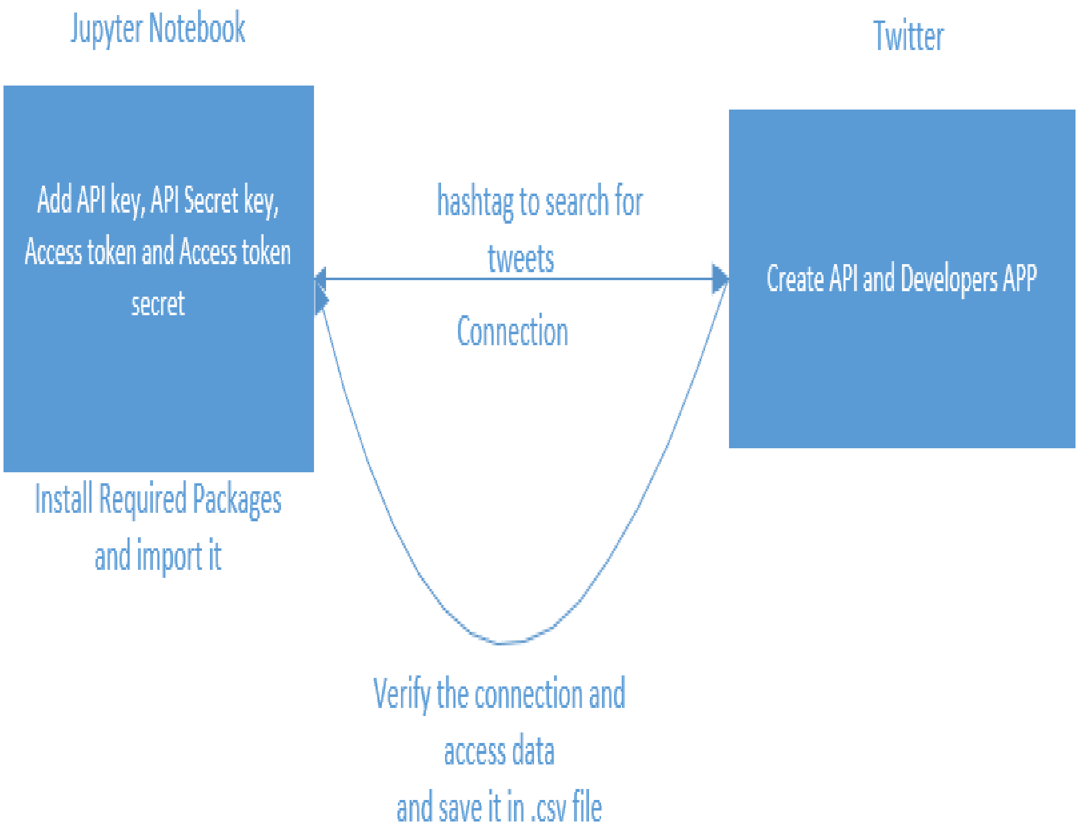


Fig. 3.1 :- Design Diagram of Project

This is the overall design diagram which shows how each phases are related to each other's outcome.

3.1. Data Collection

The design diagram for data collection can be shown as below:-



Jupyter Notebook
Twitter
Add API key, API Secret key, Access token and Access token secret
hashtag to search for tweets
Connection
Create API and Developers APP
Install Required Packages and import it
Verify the connection and access data and save it in .csv file

Fig. 3.1.1:- Design Diagram for Data Collection

The process of data collection can be explained as below:-

1. First, create a twitter account.
2. Create developers account for twitter.
3. Create an app on a developer account.
4. Save required signature keys from twitter developers account in .txt file .
5. Start Jupyter Notebook .
6. Create a workspace for python in Jupyter notebook .
7. Import required libraries.

8. Write the required codes.
9. Give those signature keys as input .
10. Provide certain hashtag to scrap data .

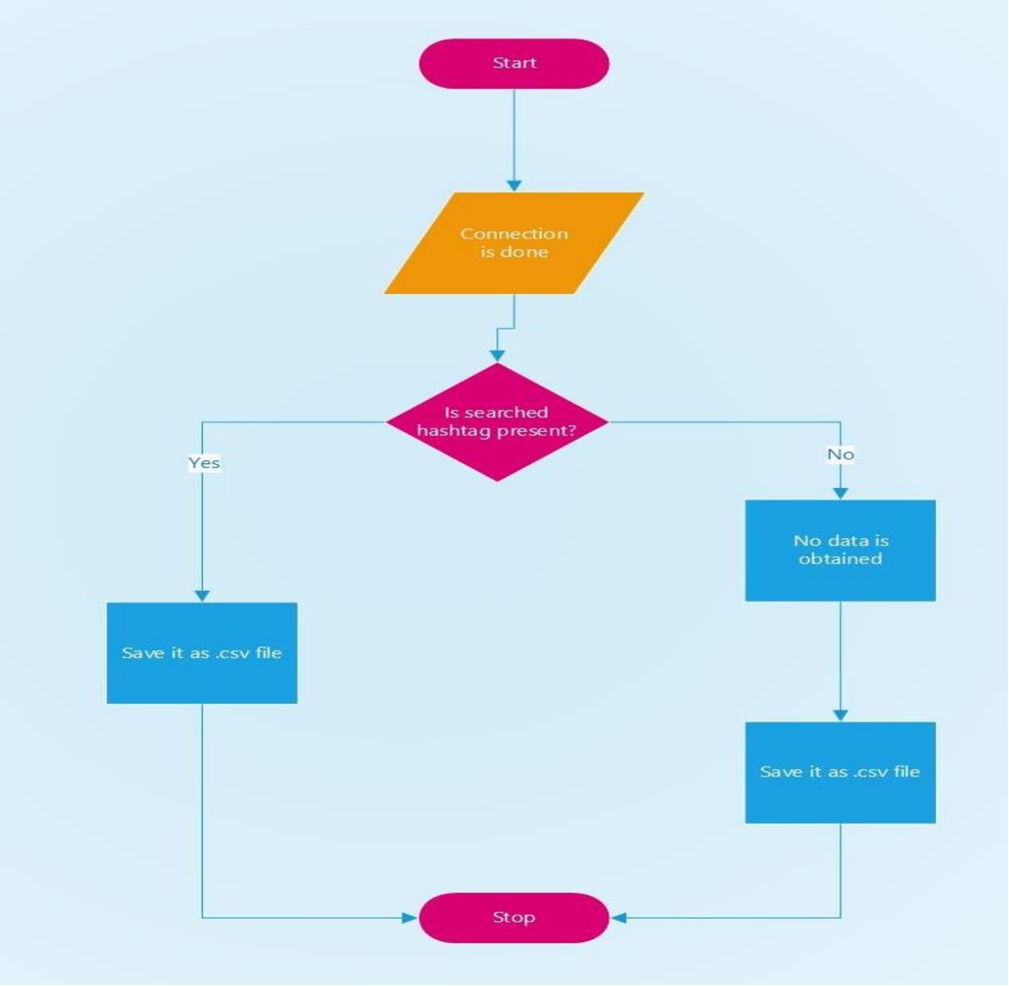


Fig. 3.1.2 :- Flow diagram for data collection

The simple flow diagram for data collection can be shown above . The process can be explained as:-

i. The connection between Twitter and jupyter notebook is established.

ii. Used certain hashtags to search in twitter.

For e.g:- #ronaldo, #messi etc.

iii. If the searched hashtags is present than data is stored as .csv file.

iv. If not it pops up the message as:- the searched hashtag is not present.

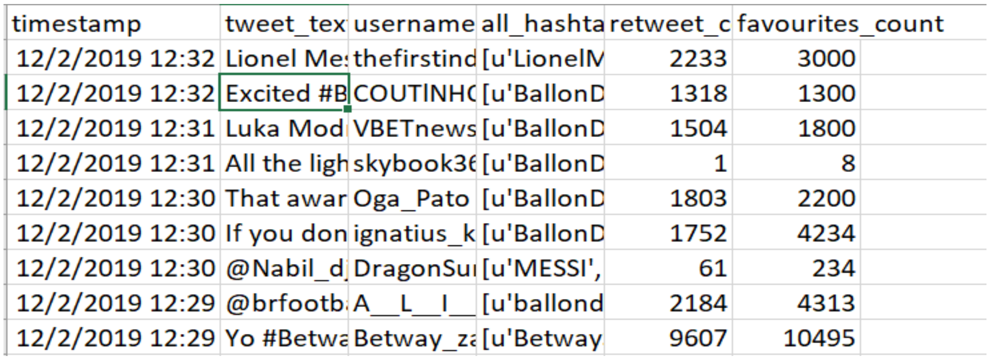
After this approach the data is collected. The tweets before the award and after the award is collected. The hashtags used are:

Table I Tweets before and after award with total number of tweets harvested

Before Award	**After Award**
#ballandor (top 100 tweets)	#ballandor (top 100 tweets)
#messi (top 100 tweets)	#messi (top 100 tweets)
#ronaldo (top 100 tweets)	#ronaldo (top 100 tweets)
#vandijk (top 100 tweets)	#vandijk (top 100 tweets)

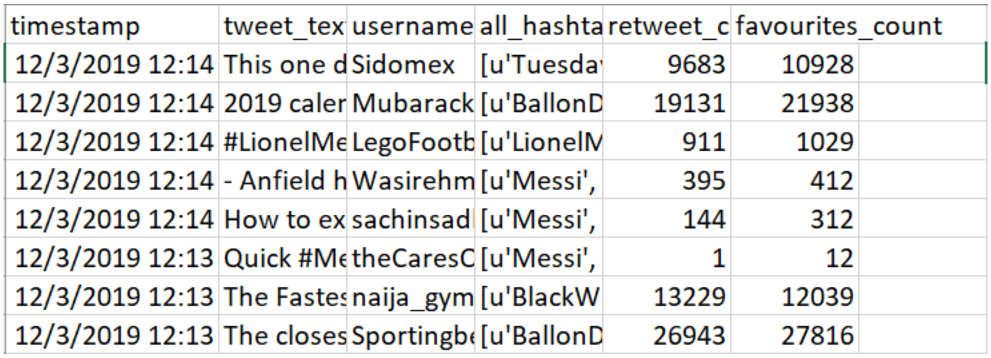
The screenshot of used .csv file of before award and after award can be shown as belows:-

Tweets associated with #ballandor before award



timestamp	tweet_tex	username	all_hashta	retweet_c	favourites_count
12/2/2019 12:32	Lionel Me	thefirstind	[u'LionelM	2233	3000
12/2/2019 12:32	Excited #B	COUTINH	[u'BallonD	1318	1300
12/2/2019 12:31	Luka Mod	VBETnews	[u'BallonD	1504	1800
12/2/2019 12:31	All the ligh	skybook36	[u'BallonD	1	8
12/2/2019 12:30	That awar	Oga_Pato	[u'BallonD	1803	2200
12/2/2019 12:30	If you don	ignatius_k	[u'BallonD	1752	4234
12/2/2019 12:30	@Nabil_d	DragonSu	[u'MESSI',	61	234
12/2/2019 12:29	@brfootb	A__L__I__	[u'ballond	2184	4313
12/2/2019 12:29	Yo #Betwa	Betway_za	[u'Betway	9607	10495

Tweets associated with #ballandor after award



timestamp	tweet_tex	username	all_hashta	retweet_c	favourites_count
12/3/2019 12:14	This one d	Sidomex	[u'Tuesda	9683	10928
12/3/2019 12:14	2019 caler	Mubarack	[u'BallonD	19131	21938
12/3/2019 12:14	#LionelMe	LegoFootb	[u'LionelM	911	1029
12/3/2019 12:14	- Anfield h	Wasirehm	[u'Messi',	395	412
12/3/2019 12:14	How to ex	sachinsad	[u'Messi',	144	312
12/3/2019 12:13	Quick #Me	theCaresC	[u'Messi',	1	12
12/3/2019 12:13	The Fastes	naija_gym	[u'BlackW	13229	12039
12/3/2019 12:13	The closes	Sportingbe	[u'BallonD	26943	27816

The Ballon D'or award ceremony took place on 3rd December 2019. So we harvested the required datasets from twitter on 2nd December for before award tweets and 3rd December for after award tweets.

3.2 Analyzing Popularity Index

The design diagram for analyzing popularity index can be shown as below:-

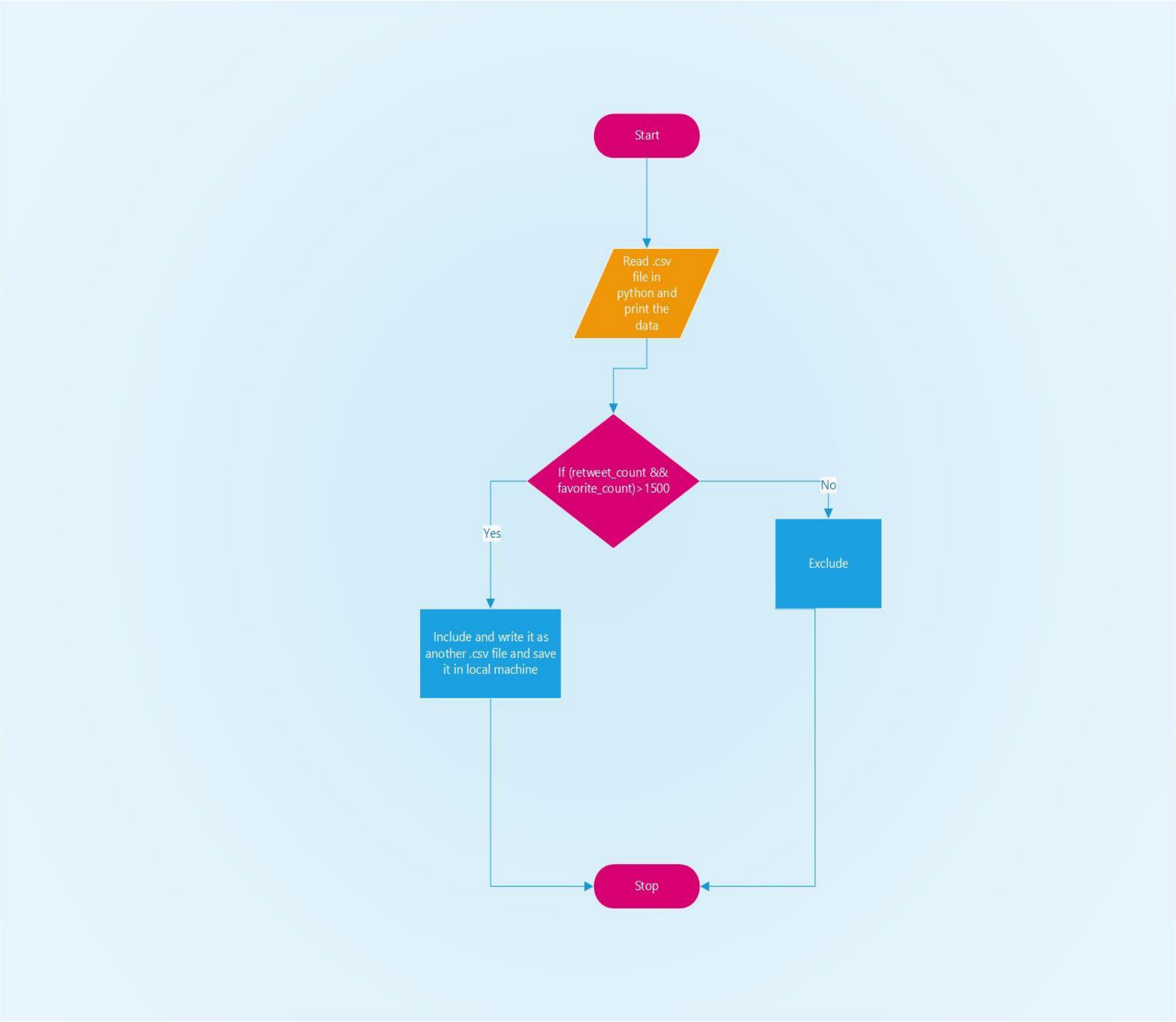

Start
Read .csv file in python and print the data
If (retweet_count && favorite_count)>1500
No
Yes
Exclude
Include and write it as another .csv file and save it in local machine
Stop

Fig. 3.2.1 :- Flow Diagram for analyzing popularity index of tweets

Algorithmic Approach

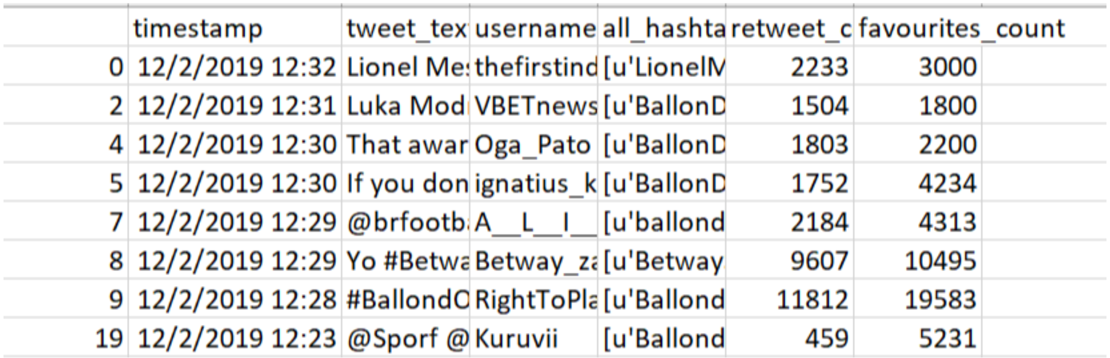
```
i.    Read the .csv file which has been harvested from the twitter.
ii.   Let retweet_count be rc and favorites_count be fc
iii.  If(rc>1500 && fc>1500)
      Print the data and write it as new csv file

      else Exclude
```

Popularity index is analyzed based on the maximum numbers of retweets and favorites count on the tweet text.

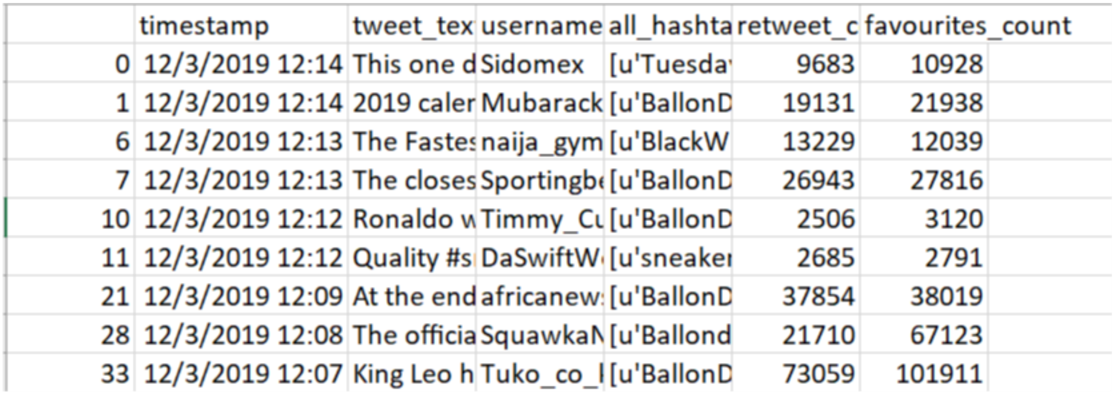
The datasets after the popularity Index is analyzed can be shown as:-

Popularity Index Analyzed for #ballandor before award



	timestamp	tweet_tex	username	all_hashta	retweet_c	favourites_count
0	12/2/2019 12:32	Lionel Me:	thefirstind	[u'LionelN	2233	3000
2	12/2/2019 12:31	Luka Mod	VBETnews	[u'BallonD	1504	1800
4	12/2/2019 12:30	That awar	Oga_Pato	[u'BallonD	1803	2200
5	12/2/2019 12:30	If you don	ignatius_k	[u'BallonD	1752	4234
7	12/2/2019 12:29	@brfootb.	A__L__I__	[u'ballond	2184	4313
8	12/2/2019 12:29	Yo #Betwa	Betway_za	[u'Betway	9607	10495
9	12/2/2019 12:28	#BallondO	RightToPla	[u'Ballond	11812	19583
19	12/2/2019 12:23	@Sporf @	Kuruvii	[u'Ballond	459	5231

Popularity Index Analyzed for #ballandor after award



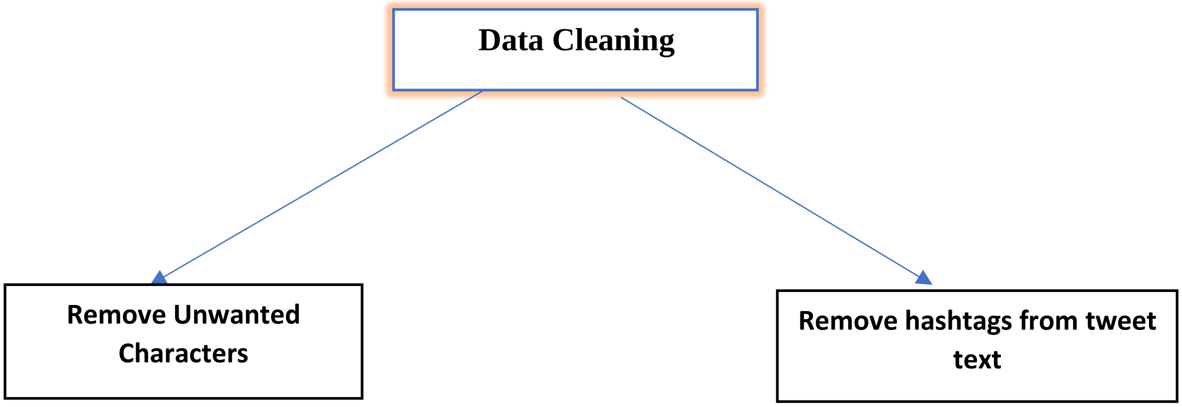
	timestamp	tweet_tex	username	all_hashta	retweet_c	favourites_count
0	12/3/2019 12:14	This one d	Sidomex	[u'Tuesda	9683	10928
1	12/3/2019 12:14	2019 caler	Mubarack	[u'BallonD	19131	21938
6	12/3/2019 12:13	The Fastes	naija_gym	[u'BlackW	13229	12039
7	12/3/2019 12:13	The closes	Sportingbe	[u'BallonD	26943	27816
10	12/3/2019 12:12	Ronaldo w	Timmy_Cu	[u'BallonD	2506	3120
11	12/3/2019 12:12	Quality #s	DaSwiftW	[u'sneaker	2685	2791
21	12/3/2019 12:09	At the end	africanew:	[u'BallonD	37854	38019
28	12/3/2019 12:08	The officia	SquawkaN	[u'Ballond	21710	67123
33	12/3/2019 12:07	King Leo h	Tuko_co_I	[u'BallonD	73059	101911

3.3 Data Cleaning

Data Cleaning is the next process that is been followed for this project. When the popularity index is determined the tweet text needs to be cleaned [9]. Hence this comes in handy.

In this phase we follow the following steps:-

i. Remove unwanted characters.
ii. Remove hashtags from tweet text.



The cleaning is performed on the popular tweets. This can be performed as below:-

i. If special characters like ☹, 😊 etc. are present in text.
ii. If hashtag(#) followed by text is present in tweet .
For e.g:- #happy, #sad etc.

The table for tweet text before and after cleaning can be shown as:-

Table II Tweets before and after Cleaning

Before Cleaning	**After Cleaning**
You can buy BallonDor but you cannot buy international trophy 😊 #goat #real	You can buy BallonDor but you cannot buy international trophy.

This data cleaning process helps us to get efficient result when we perform sentiment analysis on those text.

3.4. Sentiment Analysis

This paper majorly focuses on the tweets written in the English language. We are comparing the tweets related to 3 football players namely Cristiano Ronaldo, Lionel Messi, Virgil VanDijk, and the name of the award in which they have been nominated namely Ballon D'or. As shown in the above 3phases we have filtered the datasets. Now we will perform sentiment analysis on those tweet text.

We take the tweets as input and compare the text against the positive and negative bag of words that we have. Based on the presence of those words on the text we define either the tweet is positive or negative.

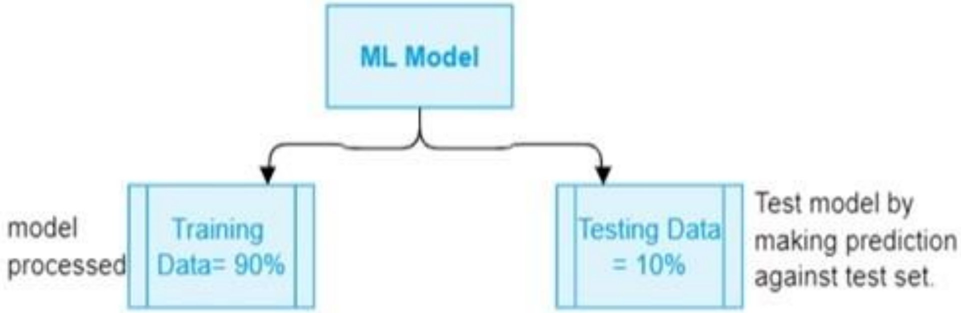
We are doing this to know how many positive and negative tweets are attached to certain hashtags before and after the award.

The positive and negative words what we have as input can be shown as below:-

Table III Positive and Negative words

Positive words	Negative words
Adorable, Adore, afford , affordable, diversified , divine etc.	abnormal, abolish, corrupt, bad , worst, abuses, accidental etc.

Performing Sentiment Analysis to this data can be started now. To do this we are using **the SVM algorithm** since it has better efficiency and provides good results.



We train the model against 90 % of those positive and negative words and test it against the 10%.

The mathematical model for sentiment analysis is divided into three phases. They are:-

i. Input
ii. Output
iii. Initialization

The model can be shown as below:-

Input :Tweet text file(text of the tweet) ‘Tt’

Output: Sentiment associated Sa= { Pt, Nt }, Strength Sn, where Pt = Positive tweet, Nt = Negative tweet

Initialization : sumPositive and sumNegative is 0 initially, Positive word Pw, Negative word Nw

sumPositive(sP) : total number of positive words in the text

sumNegative(sN) : total number of negative words in the text

Begin

```
For each Sa in Tt
    Search for Sa
    If Pw is present in Tt
        $sP = sP + Pw$
    If Nw is present in Tt
        $sN = sN + Nw$
   End If
End For
Count for number of Pw and Nw in Tt
If $sp > sN$
  $Sa= Pt$
  $Sn = sP / (sP + sN)$
Else If $sp < sN$
  $Sa= Nt$
  $Sn = sN/ (sN + sP)$
End If
```

End

The sample example for the result can be shown as belows:-

Tweet Text (Tt)= ‘ Lionel Messi wins Ballon dOr for record sixth time . But he was not that good. ’

Positive Words (Pw)= 2

Negative Words (Nw)= 4

sumPositive (sP)= 2

sumNegative (sN) = 4

Hence, sN > sP (where 4 > 2)

Sa = Nt

So, the tweet is negative tweet.

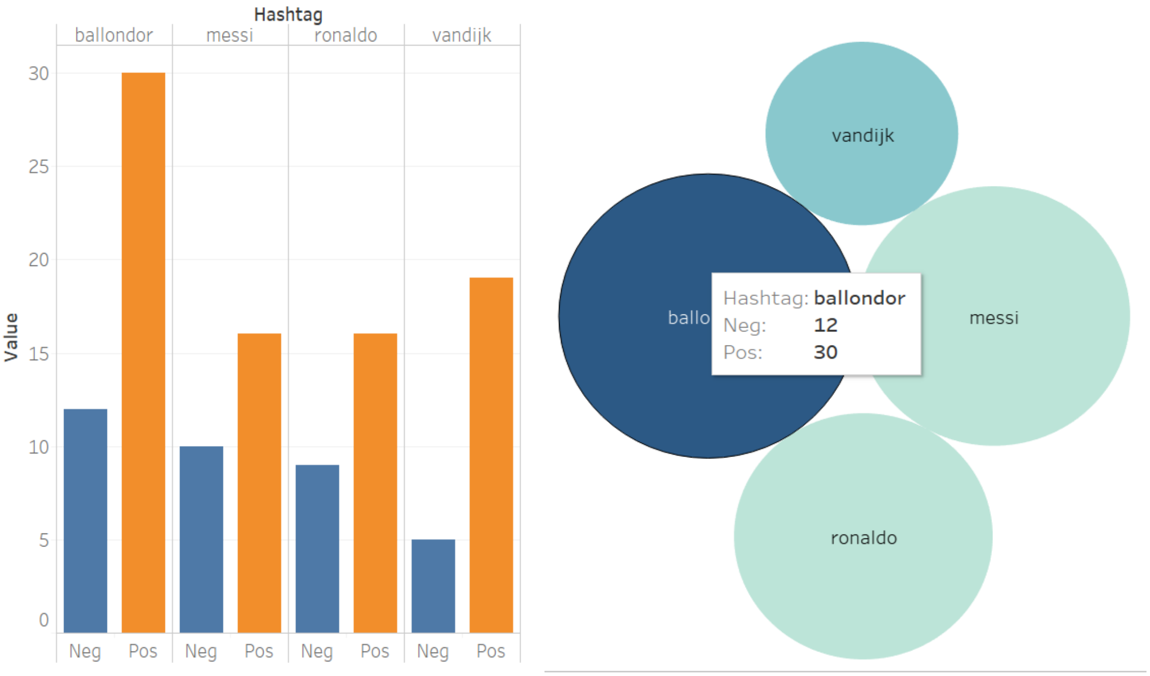
The final result of the sentiment analysis for the prioritized tweets will be shown in the results.

3.5. Data Visualization

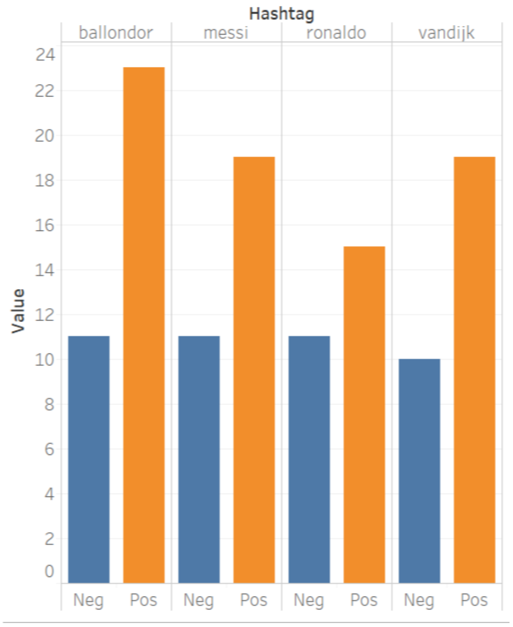
Visualization is done to know the meaningful insights into data. Visualization makes data clearer and understandable [10]. Here in this after we performed the sentiment analysis for the prioritized text we visualize the data. We visualize the data to understand the number of positive and negative tweets before and after the awards for those hashtags.

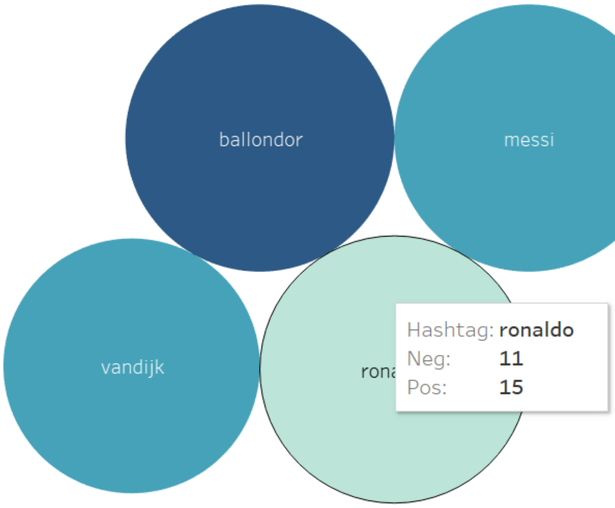
The visualization of the data for sentiments associated on tweets before and after the award can be shown as:-

Before Award



After Award





4. Results and Discussion

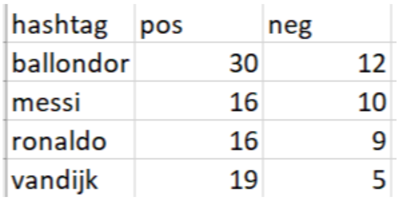
This paper focuses on analyzing or filtering popular tweets and finding the sentiment associated with those tweets. We completed all five phases and got the result too.
The result is divided into two categories:-

1. **Analyzing Popularity Index:-** As explained in the above phases the popularity index on tweets is analyzed against two parameters that are,
   i. If retweet count and the favorite count is greater than 1500
   this procedure is done we successfully filtered the tweets and only has the popular tweets.

2. **Sentimental Analysis of popular tweets:-** The popular tweet's text is used to analyze the sentiments associated with those tweets. We used the SVM algorithm to perform sentiment analysis. The efficiency of up to 70% is achieved due to this. We performed this procedure successfully and this is explained in detail in the above phases.

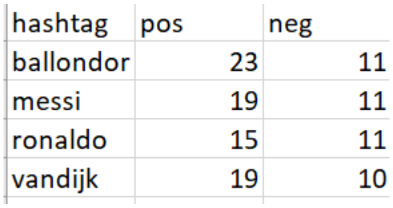
Again divide the sentiment analysis in two ways:- before award and after award.can be shown in the excel file as below:-

Before Award



hashtag	pos	neg
ballondor	30	12
messi	16	10
ronaldo	16	9
vandijk	19	5

After Award



hashtag	pos	neg
ballondor	23	11
messi	19	11
ronaldo	15	11
vandijk	19	10

As we can see in the above data, there is a difference in the sentiments before and after the award. It is due to the sentiment of their respective fans. The hashtag associated with ballandor doesn’t have much difference in terms of negative tweets but when we filtered the tweets on priority index the positive tweets are decreased.

But the same cannot be said for a hashtag associated with Messi. He was the winner of the award and it clearly shows in the tweet as well.

For the hashtag associated with Ronaldo, it is not much of a difference.

For the hashtag associated with van Dijk, it shows the difference between negative tweets on after award scenario. He was the runner-up in the award.

5. Conclusion

Sentiment analysis is the field of study where we ought to know the view and sentiment of the people. Performing sentiment analysis on popular tweets makes it more efficient. This is because when people ought to support certain tweets they retweet it and also make it as favorite. This study helps to know the views of people towards the players which have been taken on our study. And when it comes to the tweet during award ceremony than it gives more good results since it will be a trending topic.

But with pros, we expect cons as well. The problem with analyzing twitter data based on the hashtag is that when people post some tweet on twitter they use multiple hashtags that might not be related to their tweet. So, the tweet might not be related to the hashtag associated. So, due to this, there will be the circulation of false information since the hashtag used is not related to the tweet.

For our study as well we got multiple tweets that are not related to the hashtag. So, this makes the model a little inefficient when it comes to performing sentiments analysis on twitter data.

5. Future Work

In the future, a more efficient model can be developed which determines if the hashtag is related to the tweet or not. This helps to minimize most of the cons and sentiment analysis can be performed efficiently on twitter data.

References

[1] M. P. Wadhwa, "SOCIAL NETWORKS ANALYSIS: TRENDS, TECHNIQUES AND FUTURE PROSPECTS," *IEEE Explorer,* p. 3, 2014.

[2] M. B. Pooja Wadhwa, "SOCIAL NETWORKS ANALYSIS: TRENDS, TECHNIQUES AND FUTURE PROSPECTS," IEEE, Delhi, 2014.

[3] S. M. U. D. G. H. Y. G. Ali Husnain, "Estimating Market Trends By Clustering Social Media Reviews," *IEEE,* no. analyzing market trends by clustering, p. 6, 2018.

[4] K. K. K. P. Madhura Kaple, "Viral Marketing for Smart Cities: Influencers in Social Network Communities," *2017 IEEE Third International Conference on Big Data Computing Service and Applications,* pp. 106-110, 2017.

[5] R. S. M, "Application of Social Media as a Marketing Promotion Tool-A Review," Amrita School of Business," *IEEE Explorer,* 2017.

[6] Y. Lin, "https://www.oberlo.in/blog/twitter-statistics," Oberlo, 30 july 2019. [Online].

[7] H. C. P. B. B. O. Cecile Paris, "Exploring emotions in social media," in *2015 IEEE Conference on Collaboration and Internet Computing*, Sydney, Australia, 2015.

[8] D. R. Chandan Arora, "Sentiment analysis on twitter data," *International Research Journal of Engineering and Technology (IRJET),* vol. 04, no. 06, p. 6, 6 june 2017.

[9] D. P. S. A. Brahmananda Reddy, "Sentiment Research on Twitter Data," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 8, no. 2S11, september 2019.

[10] M. M. C. R. D. G. Renato Toasa, "Data visualization techniques for real-time information — A custom and dynamic dashboard for analyzing surveys' results," in *IEEE*, Liberia, 2018.

[11] M. A. R. J. J. J. G. S. M. H. Kirk Roberts, "EmpaTweet: Annotating and Detecting Emotions on Twitter," *IREC,* p. 8, 2012.