Sentiment Analysis on Soccer Players using Twitter Data

Report 2

Aravind Padmanabhan
Computer Science
Binghamton University
Binghamton, New York
apadman1@binghamton.edu

Ridhuvarshan Natarajan
Computer Science
Binghamton University
Binghamton, New York
rnatara2@binghamton.edu

Chethana Gopinath
Computer Science
Binghamton University
Binghamton, New York
cgopina1@binghamton.edu

INTRODUCTION

Racism has always been a major concern in soccer. It has been prevalent in the past as well as even in current times such as the recent incidents involving Pogba, Rashford etc. Players faced racism from fans in case of bad performance and even in case of good performance they faced racism from opponent fans. To analyze this, we had initially started our project by fetching tweets based on targeted player names and checking for racist content. But we did not get a considerable number of tweets that were racist, and they were mostly only articles and news links regarding the racism that happened. Hence, we changed our project but within the same domain of soccer. Our project now mainly focuses on the atmosphere in twitter during soccer matches by collecting the tweets with the help of hashtags. Hence in this project we collected the tweets in periodic intervals and stored them in the database. With the collected tweets we had planned to perform the analysis and come up with answers to our research questions that include

- Finding out the popularity of the English Premier League in other countries.
- The reaction of the fans of particular clubs(in Twitter) to the major happenings around a team, - for example, sacking of a manager, players stripped off captaincy, injury of star players during the season or transfer of players to other teams, match results between said team and an opposing team and such.
- Ranking the matches itself based on the popularity they gained.

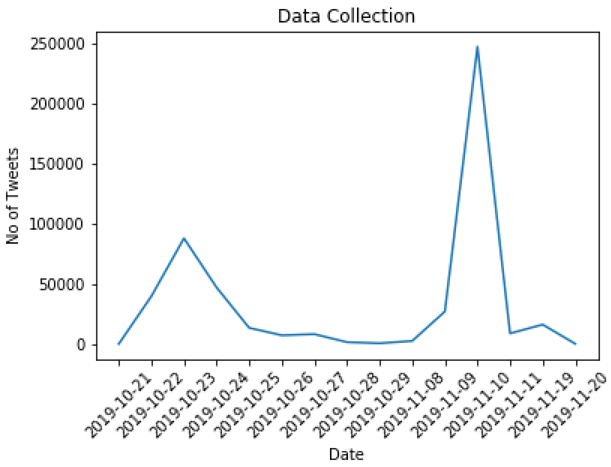
Our teams of interest include Arsenal, Manchester United, Manchester city, Chelsea, Liverpool, Tottenham Hotspur. We have followed the performance of these teams across Champions League, Europa League and Premier League.

BACKGROUND

We can always find a different perspective about performance of a team or an individual player from the reaction of the fans rather than by just looking at the standings. The tweets related to soccer by fans include discussions like benching of players, comparison between the first and second half of the match. We can compare the views of studio pundits with that of the fans using this data. The image in social media of a manger or a player can be juxtaposed with that of their real-life image to check for the similarities and differences.

DATASET DESCRIPTION

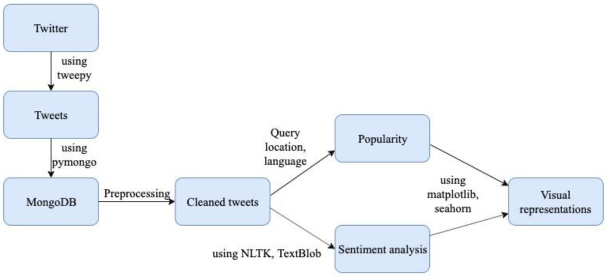
The total number of tweets that were collected over a period of two and a half months were around 500,000 with the size of around 3 giga bytes. This had to be filtered out and therefore we removed all the tweets that were in languages other than English. Once this was done, the number of tweets were around 370,000. This had to be done because we had done sentiment analysis on the collected tweets using the ‘TextBlob’ python library. Although TextBlob as well as the ‘googletrans’ libraries offered translation options, they were not effective enough in getting the actual translations. Also, tweets which were retweets were also removed. After this the number of remaining tweets were around 120,000 in number. The tweets that we collected came with all the meta data. From this the only fields that we needed to use were ‘text’, ‘location’, ‘time’ and language. We had stored each collected tweet as a whole in the database (MongoDB) and fetched only the required fields from the it. One of the surprising factors we found in the dataset was, for the language field tweets in ‘undetermined’ language was more in number than English itself. This was because the tweets consisted primarily of emojis. Just like all the important happenings, hashtag misuse was present for soccer matches also. Unrelated tweets were using hashtags of popular matches to advertise their own content.



The above figure shows the number of tweets collected by dates.

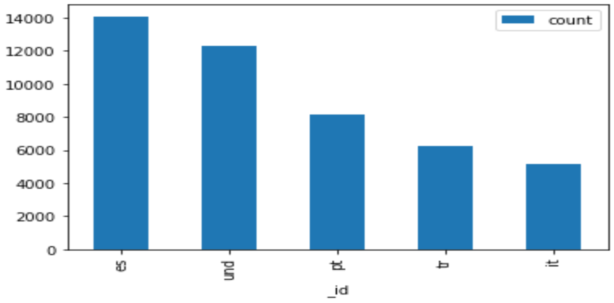
METHODOLOGY

To collect our tweets, we used a Listener class in tweepy module in Python and stored the tweets to the dataset collection using MongoDB by importing pymongo onto our Jupyter notebook. Next, we did a significant amount of preprocessing on the collected tweets - essentially a cleanup of the dataset. Here, we removed white spaces. hashtags, links, retweets, mentions and emojis within the code. We also converted all the tweet data to lowercase to make the format of the tweets simpler. Now, our data was in a suitable format to start analysis. To present our analyses visually, we made use of two modules in Python - matplotlib and seaborn. To figure out the popularity of EPL in other countries, we queried and extracted both the language and location fields from the database to trace out the origin of each tweet, thereby leading to a measure of how popular EPL was in other countries. The complication here was to find out how the users had keyed in the location, if it was 'City, Country' or vice versa. Our assumption was to interpret the text entries (for the location field) as 'City, Country'. There were also cases where there was just one entry, like the country alone was mentioned. In such cases, we assumed a null value for the other field and plotted the histograms. We also grouped hashtags together to find out the frequency of occurrences of hashtags and to see which one was used more by the users. To answer the reminder of our research questions by performing sentiment analysis of the tweets collected, we had initially planned to use spaCy but due to the needless complexities of its TextClassifier class, we decided to make use of the sentiment analysis methods provided within TextBlob and NLTK for deriving the required results. Using nltk.sentiment.vader, the positive, negative and neutral polarities of each tweet are calculated with positive being greater than 0, neutral being 0 and a negative polarity would be any value lesser than 0. The compound polarity is calculated using the above three polarities for each tweet as shown in the data frame. In addition to NLTK, TextBlob is another NLP library, which we utilized to compare sentiments between teams and people, histograms are plotted for the same. We also used TextBlob to find out positive and negative count of tweets between managers of clubs.



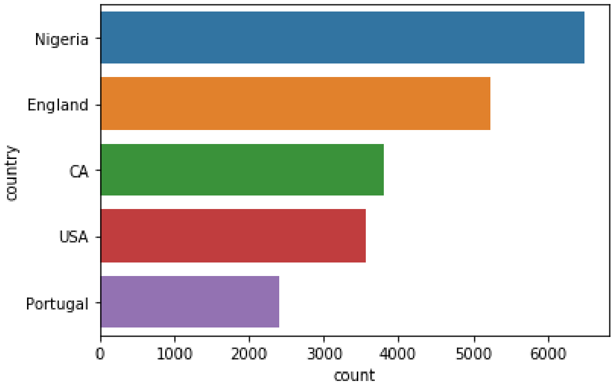
RESULTS

At first, we analyzed the popularity of the English Premier League in twitter user who use languages other than English and we got the following result



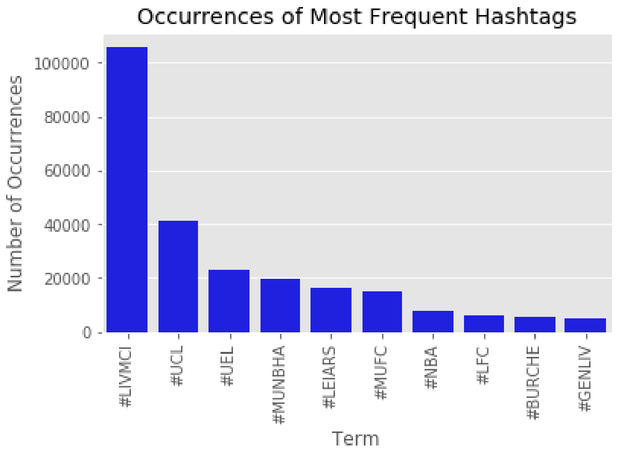
As mentioned earlier the tweets containing emojis were significant in number and this is denoted by the 'undetermined' column. Another interesting factor was finding Turkish in this list.

Next we analysed the popularity of the English Premier League by location and we arrived with the following result.



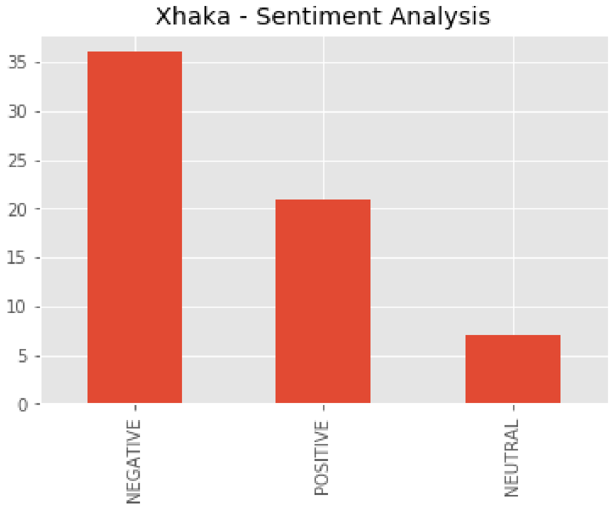
The surprising find in this analysis was Nigeria being on top of England itself! Another factor in this graph was the location CA. Since location is a free text, we had taken only countries and filtered the cities in python by splitting the words. But CA is ambiguous as it can mean both California and Canada. In case if it was California, then USA would be in the first place.

Our next analysis was finding out the most used hashtag among our list of hashtags. That yielded the following result



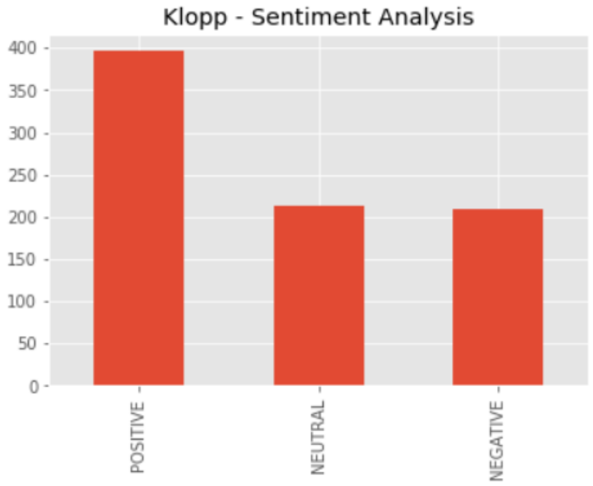
This clearly shows that #LIVMCI hashtag clearly dominated the rest. The explanation for this is that during that period, these teams were competing for the top spot. The second hashtag #UCL is for the Champions League which is being followed even more than the Premier League, but users who tweet about the Premier league, only include the match hashtags and not the #EPL hashtag itself. This could prove to be an explanation as to absence of #EPL in the top 10 hashtags.

Following this our analysis was focused on sentiment analysis of the tweets by using 'nltk.sentiment.vader'. We decided to analyse the reaction of fans on the Arsenal midfielder Granit Xhaka and we arrived with the following result



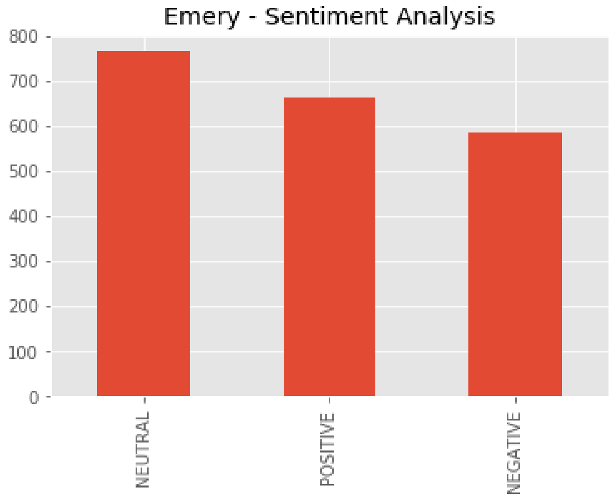
The high amount of negative opinion on Xhaka corresponds to the booing incident by the fans of his own club when he was substituted due to bad performance.

Similarly we did the analysis for the manager of Liverpool Jurgen Klopp. The result obtained was as follows



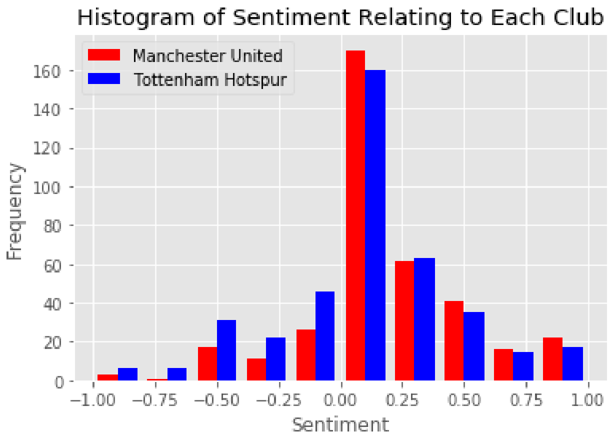
The high positive results for Klopp corresponds to his team being the first in the table.

The same process was tried for Unai Emery – the former manager of Arsenal who was sacked for the bad performance of his team. We did this expecting a high amount of negative opinion.

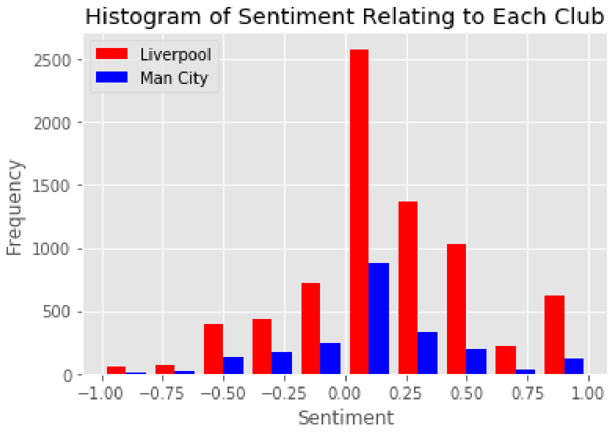


The result obtained did not confirm with our hypothesis as neutral was the highest. So we looked inside some of the tweets that were classified as neutral and found that a large number of them were opinions against Emery. This proved that the NLTK library was not 100% efficient in the sentiment analysis.

So far we had analysed only one player or team using the NLTK library. The following analysis was done using the TextBlob library for comparing two different teams at the same time. The following result was obtained for Manchester United and Tottenham Hotspur.

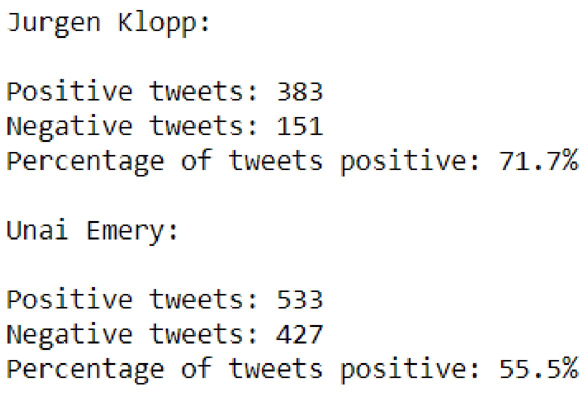


Similar experiment was performed for Manchester City and Liverpool to get the following result



At a first glance the sentiment of both positive and negative for Liverpool being higher than Manchester city was meaningless. But taking the fan count into account, it made sense as fans who follow Liverpool far outweigh the fans who follow Manchester City.

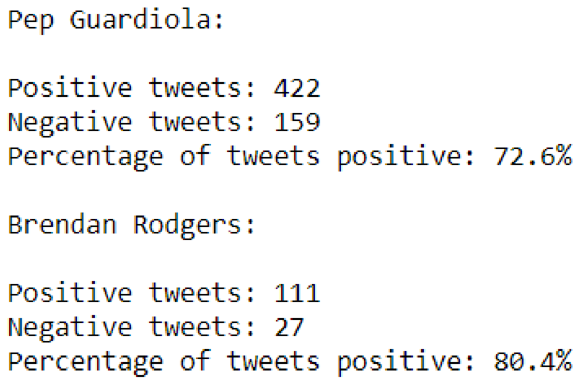
We then used ‘TextBlob’ for analysing the positive and negative tweets of two managers and compare them. The results for Klopp and Emery are as follows



```
Jurgen Klopp:

Positive tweets: 383
Negative tweets: 151
Percentage of tweets positive: 71.7%

Unai Emery:

Positive tweets: 533
Negative tweets: 427
Percentage of tweets positive: 55.5%
```

The above figure corresponds to the ranks of the teams that these managers were handling and the high amount of negative tweets for Emery corresponds to his sacking from Arsenal

A similar experiment was carried for Pep Guardiola and Brendan Rodgers and the following result was obtained



```
Pep Guardiola:

Positive tweets: 422
Negative tweets: 159
Percentage of tweets positive: 72.6%

Brendan Rodgers:

Positive tweets: 111
Negative tweets: 27
Percentage of tweets positive: 80.4%
```

Though Guardiola won the previous seasons, Rodgers has a higher number of positive tweets because the standing for his team was better at that point

DISCUSSION AND CONCLUSION

From the results obtained, we can infer that they mostly corresponded with our hypothesis. The high amount of negative sentiment for Emery and Xhaka proved that the reaction on social media are in correspondence with the actual events. Some shortcomings of this paper were that the data collected was not consistent and was non periodic. Another one was skipping non English tweets as we faced the challenge of translation. Also performance of soccer clubs in other leagues that do not affect the rankings of that league could also elicit reaction from fans and these could impact the inference we formed between the ranking of teams and reaction from fans. Future developments of this project could implement analysing the sentiment of fans during matches itself and comparing them with factors like first half vs second half, after a goal is scored or a controversial decision is given etc

REFERENCES

http://docs.tweepy.org/en/latest/

http://www.nltk.org/howto/sentiment.html

https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

https://pypi.org/project/pymongo/

https://www.premierleague.com/tables

https://stackoverflow.com/questions/33404752/removing-emojis-from-a-string-in-python