Football TweetsSocial Network Analysis

Henok Birru

Abstract

According to Forbes' list of the top worldwide applications by monthly active users in 2022, Twitter is rated eighth[1]. Additionally, twitter.com had 2.4 billion sessions in 2021, with 620 million of those being unique[2]. This indicates that users visit the Twitter website frequently. One of the newer domains where data-driven insights are used to create strategy is sports analytics. There are more football-related tweets than any other sport; 70% of twitter users claimed to follow, watch, or be interested in football on a regular basis[2]. Of all the football leagues, the Premier League has the largest global broadcast audience. In this report, Twitter data for the 2022 English Premier League(EPL) is analysed using machine learning algorithms to determine the sentiment of individuals around the world. By gathering and analysing the tweets, sentiment polarity was determined. The report also contains network analysis for the top mentioned individuals, Named Entity Recognition(NER), and team fan base country distribution. The average sentiment scores for each of the top 5 EPL teams this season were calculated to provide a response to the question of how people feel about them.

Introduction

According to a SPORT+MARKT survey, more than 1.4 billion people identify as fans of a certain Premier League club worldwide[3]. Football is a business that is fueled by feelings. Sports fans frequently express their thoughts and emotions about what is happening at various matches through tweets. Big Data analysis and machine learning techniques have helped football analysis more and more in recent years, particularly in an effort to analyse tactical behaviour and identify strategy-enhancing tactics.

In this study, sentiment and network analysis are conducted on Twitter, one of the social networks most popular with football fans. Sentiment analysis is the automatic technique of identifying the user's emotions from their written content by analysing unstructured data and creating a model to derive knowledge from it. Making sentiment analysis on twitter messages is not that straightforward as others' content because we have some short texts, and informal ways of communicating. Although there are other free machine-learning models, VADER from the nltk library is utilised in this experiment. VADER(Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media[4].

Named Entity Recognition (NER), in addition to sentiment analysis, is also included in this work. NER is a technique for figuring out what entities exist in a text document. The top five entities classified as organisations and the top ten entities identified as people in the tweets are shown in this report. Different standard libraries, such as Stanford Named Entity Recognizer(SNER), can be used to apply NER to a given text. The en_core_web_lg from the Spacy library is utilised in our implementation. Spacy is a popular Natural Language Processing(NLP) library which has 84 trained pipelines for 24 languages. The en_core_web_lg is an English pipeline optimised for CPU and free to use.

By using the google maps api service, we made analysis on the locations of the football fans worldwide. The data fetched from twitter has the address in different formats, to provide better analysis the locations are grouped by countries. To achieve this, we took the advantage of the google geocode API service.

Finally, this study implements partial network graph analysis for the top mentioned users. The Networkx python library is used to create the network graph and determine the nodes' degrees.

Data and Methods

Data

50,000 tweets from 2022-08-06 to 2022-10-30 that contain the hashtag "#premierleague" make up the bulk of the collection. The 2022-08-06 date is selected because it is the start date for the current EPL season. Only English-language tweets were allowed, and in our search, links and replies were excluded. To access this amount of twitter data within the specified period of time the snscrape library was used. The data included the following information for each tweet: URL, date, content, id, user, replycount, retweetcount, hashtags, mentionedusers and so on.

Later in our work, the follower ids of the top mentioned users from the primary data set were added to do the network analysis part of the study. Since the amount of data required for this part was not large, the Twitter API was utilised via the developer elevated access. To interact with the API we used the Tweepy python library.

Data Cleaning

For our analysis, the tweets that are not useful for our next tasks were removed. Some of the tweets contain just predictions and some of them are ticket sales for different matches and others are completely different sports such as cricket. Additionally, only some of the information from the primary data are needed in our experiment, thus we removed most of the columns. Although the user information is included in the data, it was in one column with python dictionary data structure thus, user name, user display name and user location were extracted from the user column and placed on their own.

After the initial data cleaning, the shape of the data is reduced to (48841, 6), rows and columns respectively. The final columns which were kept for the next task are: date, content, hashtags, username, userdisplayname, and userlocation. The next data cleaning process was applied to the content column. Twitter data contains different emojis, hashtags, links and

mentions. These parts of the tweets were cut from each tweet since they didn't contribute anything to our sentiment analysis.

The final data cleaning process was applied before calculating the sentiment score. In this step, using the NLTK TweetTokenizer module, the tweets are tokenized. Tokenization is used in natural language processing to split paragraphs and sentences into smaller units that can be more easily assigned meaning. Following the tokenization process, the stopwords were removed from the tweets. Stopwords are the words in any language which do not add much meaning to a sentence. And finally punctuation and digits were eliminated for the same reason as the stopwords. The final cleaned tweets are stored in another column called cleaned_content.

Calculating Sentiment Score

The method of "computationally" assessing whether a piece of text is good, negative, or neutral is known as sentiment analysis. To calculate the sentiment score of the cleaned tweets vader_lexicon module is loaded from the nltk library in our experiment. VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.

VADER gives us the negative, positive, neutral and the compound score. The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalised between -1 and +1. If the compound sentiment score of a tweet is less than 0 then it is labelled as negative sentiment. If the compound score is greater than 0 then it is labelled as positive sentiment and the tweets with zero compound score value are neutral.

Named Entity Recognition(NER)

NER is a process of identifying different entities in a given text. NER models can find different entities that are present in the text such as persons, dates, organisations, and locations. In our work, to extract entities from the tweets, first the en_core_web_lg module was loaded from the spacy library. The Spacy library provides four different trained pipelines for the English language: en_core_web_sm, en_core_web_md, en_core_web_lg, and en_core_web_trf. The en_core_web_lg is optimised for CPU and It has tok2vec, tagger, parser, senter, ner, attribute_ruler, lemmatizer components. The NER data is extracted to another dataframe and analysed in our work, the result will be discussed in the Analysis section.

Network Analysis

Investigating social structures using networks and graph theory is known as social network analysis. In our work, we built a network graph for the top mentioned users from the initial dataset. First the user ids of the top 11 most mentioned users were extracted and filtered. Using the selected user ids their followers id were collected. To access the followers data, the Tweepy library was used. The Twitter API provides access for developers to access the data and to access their methods Tweepy was installed. The number of followers for the users were restricted to 5000. If a user has more than

5000 followers then only the first 5000 followers were taken.

Because of the limits the Twitter API set, after we fetched the data, a data frame with source and target columns were created and stored as a csv file. The source column contains the initial users(top mentioned users) and the target column contains the user id of the follower for the given source user id. To build the network the Networkx library was used. The network was created using the from_pandas_edgelist method by providing the created dataframe.

Result Discussion

The purpose of this work is to answer a few questions: What are the people's feelings related to the English premier league?What are the most famous entities in the English premier league? Which countries have more premier league fans?. In this section, the analysis derived from the tweets are discussed.

The first analysis was about the sentiment of the tweets related to english premier league.Although the accuracy of the model specially on the twitter data which are more informal should be taken into consideration, the overall sentiment score indicates there are more positive tweets than both negative and neutral. Figure 1, depicts that 47% of the total tweets have a positive intention and 31% of the tweets are neutral whereas the 22% are negative sentiment tweets.

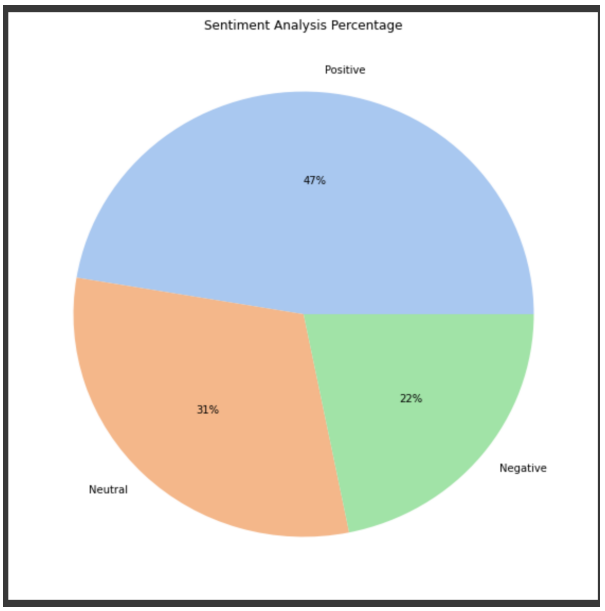

Figure 1: Sentiment Analysis Percentage

"i find it so satisfying not having to scroll an inch when looking at the table and every stat are always their its a wonderful life for newcastle fans right now" is one of the tweets which is labelled as positive sentiment. "**var is utterly killing football.any joy is put on pause which takes the moment away.** " this tweet is categorised as a negative sentiment. There might be mislabeling somewhere in our data, because social network data is not as clear as other contents and for the football sector it will be even difficult.

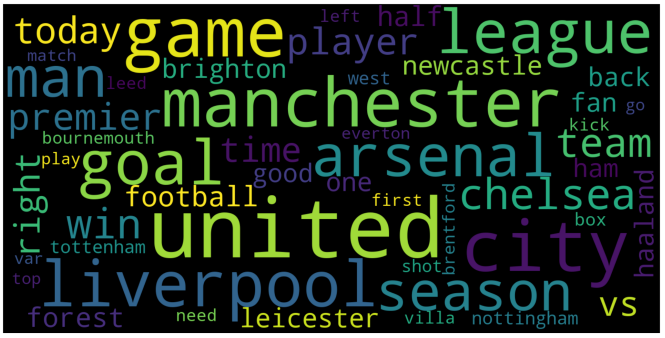

Figure 2: Word Cloud for English Premier League

As shown in Figure 2, most mentioned words in the tweets are the top team in the league; such as Manchester United, Liverpool, Arsenal, Chelsea e.t.c. On the other hand, we can see new words for example: var, worst, referee, e.t.c in Figure 3, which displays the word cloud for the negative sentiment tweets which indicates that there are people who are not happy with the VAR technology. The VAR technology supports the decision-making process of the referee, however football fans sometimes find it uninteresting because different reason, the tweet mentioned previously is one reason.

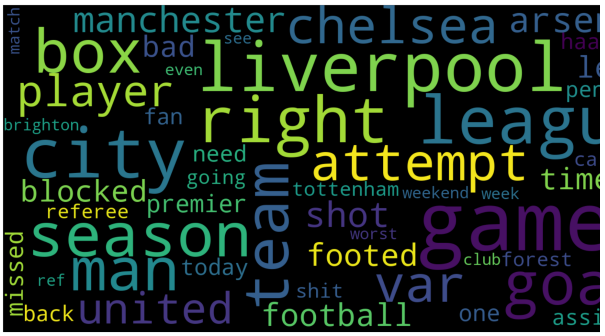

Figure 3: Word Cloud for the Negative Sentiment

The Spacy library has different built-in entities, for instance: PERSON, ORG, GPE, DATE, CARDINAL, and so on. In our case we only focused on PERSON and ORG entities. The model sometimes identifies the clubs as ORG or GPE because the name of most of the clubs are also city names. In addition, the model mis-assigns the entities as PERSON although they are club names. As we understood from our experiment building our own NER model or sentiment model would be better than to use the trained model. However, labelling our data and building the model is more costly than using already trained standard models. Thus it is a tradeoff that we should consider.

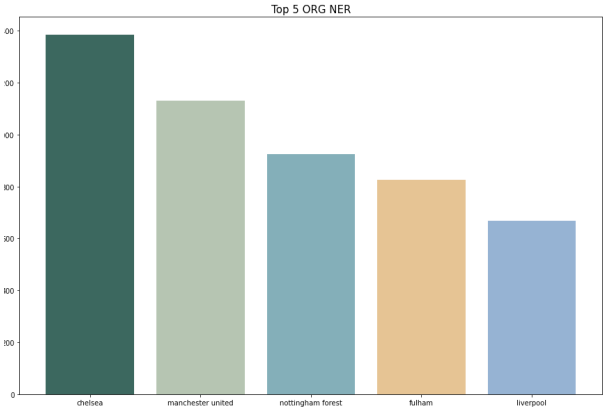

Figure 4: Top 5 ORG NER

The top 5 entities identified as ORG are Chelsea, Manchester United, Nottingham Forest, Fulham and Liverpool as shown in Figure 4. The first 10 entities identified as PERSON are displayed in Figure 5. Brighton, Manchester United, Chelsea, and Wolverhampton are also identified as PERSON, which is incorrect, however the other entities such as Ronaldo, Antony, are the correct identification.

Manchester United is one of the entities mentioned in both categories. This indicates there are many tweets related to the club and according to stadium maps this club is number one with 75,012,095 number of fans[5]. Time Series sentiment analysis was done in our work for this club for the 2022 season so far. The sample is taken by weekly interval using pandas resample method. As Figure 6 depicts the sentiment score of the club increases overtime.

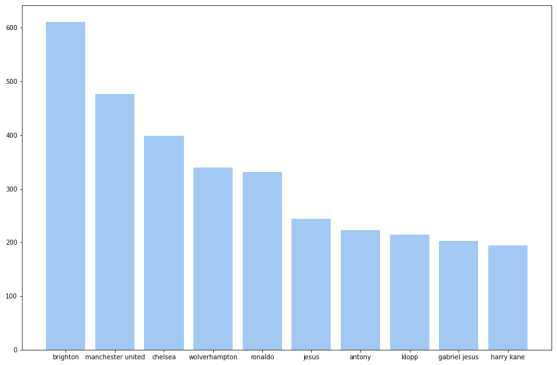

Figure 5: Top 10 PERSON NER

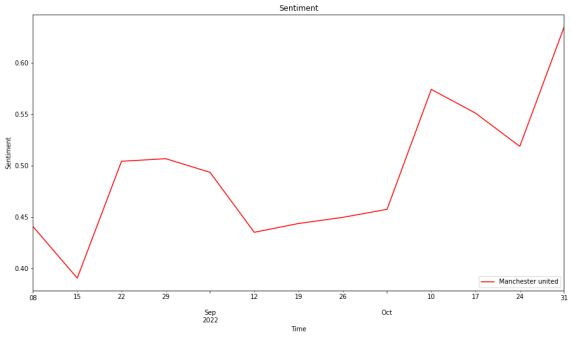

Figure 6: Manchester United Timeseries Sentiment Analysis

Additionally, following Manchester United, Chelsea, Manchester City, Liverpool and Arsenal are also mentioned in the top 5 clubs with enormous numbers of fans. The average sentiment score was computed for these teams, however the distribution of the number of tweets might have an impact, as shown in Figure 7, Manchester United has an average positive score than the other clubs.

The other interesting question that we can answer using this twitter data is which countries watch the EPL most. Although, the analysis only considers those fans who watch the football and tweet about it, thus we are not completely concluding that these are the top countries that watch the EPL.

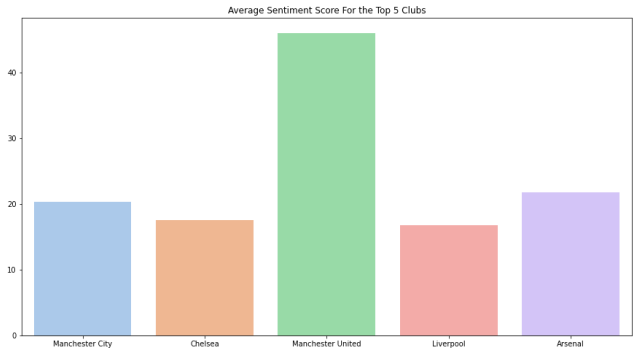

Figure 7: Average sentiment score for the top 5 English Premier League Clubs

As Figure 8 depicts, overall, the UK, USA, India, and Nigeria are the top countries which watch and tweet about EPL. Figure 9a displays the countries which tweet about Manchester United and Figure 9b depicts the countries which tweet about Arsenal, from these figures we can tell that there are more Arsenal fans in the USA than Manchester United.

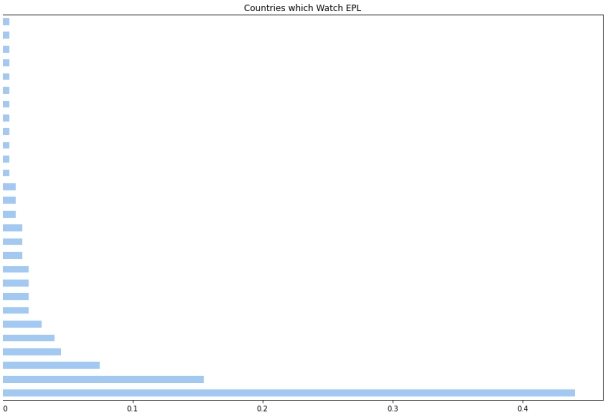

Figure 8: Countries which watch EPL

The final analysis that has been conducted in our work is the network analysis between the top 5 mentioned users from the original dataset and their followers. As shown in Figure 10, except the one node that is placed on the left side, all are concentrated on the center. The names of the selected users as a source are Manchester United, Arsenal, Premier

League, Liverpool FC, Chelsea FC, Manchester City, Erling Haaland, Fantasy Premier League, Tottenham Hotspur, Nottingham Forest FC, Cristiano Ronaldo. These users were mentioned most on the original twitter data, also from Figure 10, it can be concluded that they have many followers since the graph is concentrated on the center. The node at the left side with less number of followers is Nottingham Forest FC with a user id 39754653, which has almost 200% less followers than Cristiano Ronaldo.

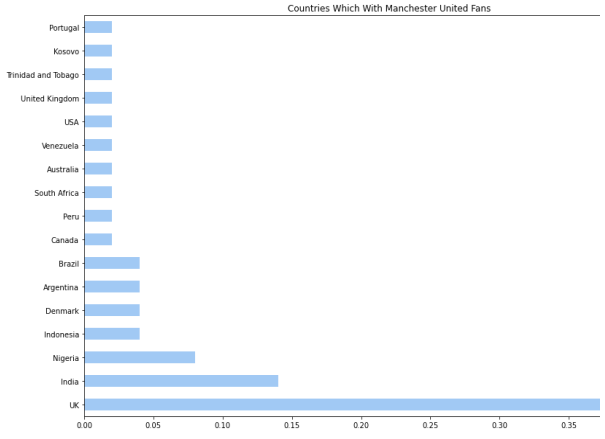

Figure 9a: Countries Manchester United Fans

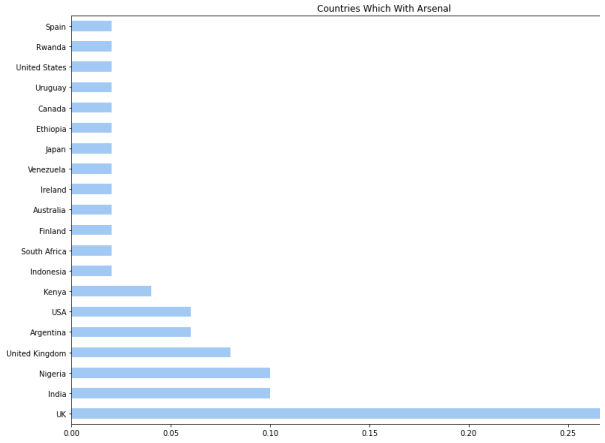

Figure 9b: Countries with Arsenal Fans

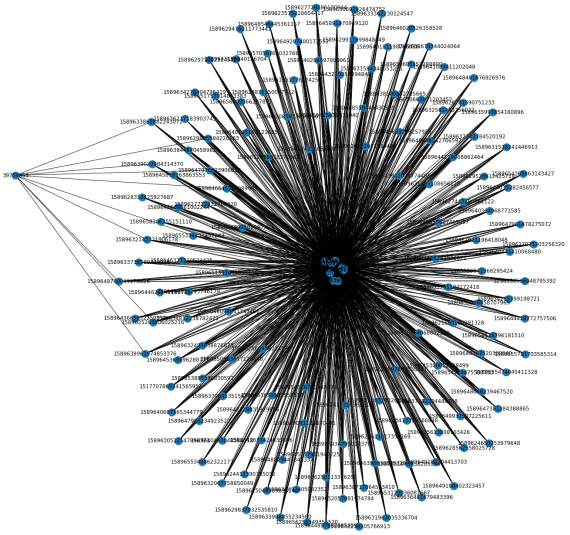

Figure 10: Users Network graph

Conclusion

In this paper, we picked one of the most popular sports in the world, Football and one of the most visited platforms in the world, Twitter and drew an analysis for the English Premier League. According to our analysis there are more positive sentiment tweets related to EPL than the other sentiment modes. Average sentiment score calculated and Manchester United shows better values than the other top 5 teams. Users-followers network analysis was also conducted in our experiment and we saw that although Nottingham Forest FC is listed as the top mentioned user, it has less number of followers than the other entities.

The challenges we faced while doing this work was first getting all the data without the limitation of any time period. The twitter API with elevated developer access can give us a limited number of tweets and they are only one

week old, however for our task we need the data starting from the first day of the 2022 EPL season. To overcome this challenge, other libraries which can give us the data without any limitation were searched and one of the libraries was Twint and the other is snscrape. The twint library is not maintained currently and it has few bugs on the current master branch, although it might be possible to fix and use twint, there is another library that can provide us the data quickly without doing much which is snscarpe, thus we use snscrape to fetch the first data for our analysis task. However, we also used the Twitter API to get the followers id for the network analysis part of this work.

The other challenge was to pick the right sentiment analysis and NER model. Although there are different trained models available for free, most of them are not good for twitter data and specially for football related tweets. While we were exploring different tools for the sentiment analysis, twitter-roberta model which uses the RoBERTa model introduced by facebook researchers was one of the tool we wanted to use but it is not suitable for our situation because it would take time to determine the sentiment, however since it is with almost 58M tweets that would perform better. Therefore, VADER is used because first it is good for social network data and also quickly calculates the sentiment score.

We saw that there is no public dataset to build a sentiment analysis or NER model from football-related tweets and since different analyses can be done about football that could help coaches, match analysts, broadcasters and other individuals we suggest working on this area.

References

[1] John Koetsier, “Top Apps Of 2022 By Installs, Spend, And Active Users: Report”, May. 2022

[2] Claire Beveridge, “33 Twitter Stats That Matter to Marketers in 2022”, March.2022

[3] SportingIndex, “FOOTBALL TV AUDIENCE FIGURES – WHO IS THE PREMIER LEAGUE’S MOST WATCHED CLUB?”, October.2021

[4] Hutto, C.J. & Gilbert, E.E., “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI,”, June 2014.

[5] Stadium maps, “English Premier League Teams Popularity”, June.2022