International Journal of Engineering &Technology, *7 (3.12) (2018) 452-455*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analysis of Football Data on Twitter for Popularity Mapping and Transfer Predictions

Raghav Murali[1*], Shikhar Shrivastava[2], Sornalakshmi Krishnan[3]

[1, 2, 3]Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai,
**Corresponding Author Email: [1]raghavmurali95@gmail.com, [2]shikhar2121995@gmail.com,*
[3]sornalakshmi.k@ktr.srmuniv.ac.in

Abstract

Twitter has gathered a reputation of being a reliable source for predictive modeling on various domains such as flu trends, sports data, political trends etc. Football is widely considered to be the most popular sport in the world. We introduce a novel approach of analyzing the tweets collected over a period of time which are related to football. We present a visualized world map with the high density areas indicating the parts of the world where football is most popular. Further, we incorporate the fan opinions of popular football players by analyzing their tweets individually. This ultimately leads to prediction of player movements from his current club to another club in the transfer window. Hence, this model helps in identifying the popularity trends in football around the world and also increases the role fans play within the club they support.

Keywords: *Tweepy, TextBlob, Sentiment Analysis, Text Classification, SVM Algorithm, Tableau*

1. Introduction

Twitter was first launched in July 2006 as one of the first websites based on social networking and micro-blogging. Over the last few years, Twitter has become a major player in the micro-blogging industry, with millions of users logging into their accounts every day. As of March 2016, the number of twitter accounts reached the 500 million mark. Twitter has been used for analyzing or predicting a variety of events such as election results, box office revenues of movies in advance of their release and the spread of diseases. All these examples are proof that Twitter data can be considered a reliable source for a wide range of studies across many domains. Hence, twitter data will act as a base for our analysis on football trends. Football is widely popular sport which has professional leagues in major countries around the world. We start by gathering all the raw tweets from around the world by means of a Python package. These tweets can then further be filtered so that they are relevant only to football. The geographical locations of these tweets are then used to plot a world map indicating the high density areas where football is most popular. Further, we try to address the highly prevalent problem of low fan participation in the way their favorite club is run. Each tweet which is collected needs to be analyzed individually. The transfer window in football is one of the most anticipated and exciting periods of the season where a large number of high profile players move to another club. Our aim in this project is to predict the transfers which are most likely to occur during this period by using the data collected from the fans‟ tweets. Sentiment Analysis is used as a basis for this module where each tweets is categorized according to its sentiment. The prediction is done using a popular text classification algorithm (Support Vector Machine) which uses the tweets as the training and test data to predict the outcome.

2. Literature Review

There is a great deal of work that has been done in analysis of twitter data in the sports field. This literature survey presents some examples of the research work in this field. Byungho min et.al (Jinhyuk Kim, Chongyoun Choe, Hyeonsang Eom, R.I.Mckay) presented a framework for sports prediction using Bayesian interface and rule-based reasoning, together with an in-game time-series approach. Joel Brooks et.al (Matthew Kerr, John Guttag) applied a supervised machine learning model to pass locations in event data from the 2013-14 La Liga season for developing a data-driven player ranking system. Keisuke Doman et.al (Tashi Tomita, Ichiro Ide, Daisuke Deguchi, Hiroshi Murase) presented a twitter based event detection method based on "Twitter Enthusiasm Degree (TED)" toward generating a highlight video of a sports game. Mitsumasa Kubo et.al (Ryohei Sasano, Hiroya Takamura, Manabu Okumura) presented a method of generating live sports updates from Twitter posts on an event. This method selects descriptive and prompt tweets that are posted within a short time after Analysis of Football data on Twitter for popularity mapping and transfer predictions important sub events by exploiting users called „good reporters‟, who promptly explain what is happening at each moment throughout the event. Glenn Healy analyzed models for predicting the probability of a strikeout for a batter/pitcher matchup in baseball using player descriptors that can be estimated accurately from small samples. This is done by using a log5 model which has been used extensively for describing matchups in sports. Carson K.Leung and Kyle W.Joseph presented a sports data mining approach which helps discover interesting knowledge and predict outcomes of sports games such as college football by using a probabilistic approach.

Markus Lochtefeld et.al (Christian Jackel, Antonio Kruger) presented „TwitSoccer‟, a real-time live soccer score ticket crowd sourced from tweets. „TwitSoccer‟ combines pattern based approach to process the tweets with a soccer knowledge database.

3. Proposed System

The proposed methodology is divided into 5 main modules:-

1. Data pool module
2. Parse and Filter Module
3. Visualization Module
4. Analysis Module
5. Prediction Module

A. Data Pool Module

The first module is the Data pool module which acts as the base for our study. It generates the data from twitter by streaming live user tweets over a period of time. The first step in order to access the twitter data is to create our own twitter application. After this is done, we will be presented with some authentication keys (Consumer key, Consumer Secret, Access Token and Access Token Secret). These keys are used to authenticate our application and are specified as fields in our code. The data collection is done by means of a Python package called „Tweepy‟. By importing „tweepy‟ in our code, we gain access to the twitter database and can stream live user data. The data obtained in this case is in a raw format and can be used to do further analysis. The tweepy package allows the user to access a list of streaming properties such as user_name, keyword track, geo location, geo coordinates etc. It is important to note that this data pool must be cleared and refreshed at regular intervals to prevent any redundancy or loss of data. Only a portion of twitter data is accessible in any period of time (1%, 10% etc.). So in order for the analysis to take place, it is important to organize the data into various data sets so that it is easier to access the data in the future. The important thing to be noted is that the data extracted in this module is completely untouched. In many cases, it might be possible that large amounts of data are irrelevant to our cause. Hence, here is where the functionality of the next module comes into play.

B. Parse and Filter Module

The parse and filter module is mainly used to filter the raw data collected previously in order to ensure that only the relevant data to our study is filtered out. The data is received from the data pool module and the format of the data is analysed. The data is then parsed into UTF-8 format which makes it easier to process it further. Then, a list of filters is applied on the data which removes the unwanted data and keeps only the tweets related to football. This is done by creating a dictionary of football related keywords which is fed into the code and compared with the data collected from twitter. Each and every tweet has a property which discloses its geographical location. Hence, the geographical location is used to analyse and group the tweets based on the country of origin. A normalized data count is then assigned to each country which is then used to plot the map. This helps us in identifying the high and low density areas of football-related tweets around the world.

C. Visualization Module

The main function of the visualization module is to plot the popularity trends on a world map which is indicated by the density levels. This is done by using popular software called Tableau. Tableau software is interactive data visualization software focussed on business intelligence. Tableau has a mapping functionality and is able to plot latitude and longitude coordinates. The filtered data obtained in the previous module is stored in a CSV (Comma Separated Value) file which is then imported into the Tableau workbook. The data in the CSV file is then used to plot the dot-map and density-map which indicate the popularity trends in football around the world.

D. Analysis Module

The analysis module acts as a precursor to the prediction of player transfers. There are mainly 3 phases involved in the analysis module:-

Text Pre-processing:-

It is important the twitter data to be used for prediction goes through a sequence of pre-processing steps in order to clean the data and remove any noise. The various steps involved in text pre-Processing include: -

1. Converting the whole data set into lower case.
2. Removing all the URLs from the tweets.
3. Converting all the user names to a common parameter such as „AT_USER‟.
4. Removing all the additional white spaces.
5. Replacing all the hashtags with the actual word (For eg: #word is replaced by word).
6. Removing all the stopwords (a list of words which don‟t add meaning to the tweet).
7. Tokenizing the tweets into individual tokens.

Sentiment Analysis**:-**

The next phase in the analysis module is to find out the sentiment of the tweet by using sentiment analysis. We Analysis of Football data on Twitter for popularity mapping and transfer predictions do this by using a python package called Textblob. Textblob is a simple API used for diving into common natural language processing tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, translation and more. By using Textblob, we assign a sentiment score to each tweet which is a combination of two fields- polarity and subjectivity. By analysing the sentiment score, we are able to classify the tweets as positive, negative or neutral. Further, these tweets are stored in a CSV file, along with their sentiment and are used for the prediction process. The table indicates some of the sample tweets classified on the basis of their sentiment score.

Table 1: Sentiment Analysis on Tweets

Tweet	Sentiment	Score
Allegri coming to	POSITIVE	0.7
Arsenal would be		
genuinely amazing.		
Chelsea beat Man	NEUTRAL	0
city 3-0.		
Worst Liverpool	NEGATIVE	-0.6
performance in a		
long long time.		

Comparison of algorithms:-

In order to determine which algorithm is best suited for text classification, we do a comparative analysis of the following 3 algorithms:-

1. Support Vector Machine
2. Naïve Bayes Classifier
3. Decision Tree Classifier

According to the, scikit‟ learn documentation, in order to implement multi class classification, LinearSVC‟ can be used in order to implement SVM. The classification report was generated when the above mentioned 3 algorithms were applied in order to conduct sentiment analysis, and yielded the following results. The parameters used are precision, recall, f1-score and Support which together contribute to determine the overall accuracy.

Table 2: Comparison of Algorithms

Sentiment	Precision	Recall	F1-	Support
			score	
NEGATIVE	0.79	0.33	0.47	228

NEUTRAL	0.69	0.91	0.79	614
POSITIVE	0.79	0.68	0.73	494
Avg/total	0.74	0.73	0.71	1336
SVM score is : 0.728293413174				
NEGATIVE	0.92	0.05	0.10	228
NEUTRAL	0.68	0.89	0.77	614
POSITIVE	0.67	0.70	0.69	494
Avg/total	0.72	0.68	0.62	1336
Naïve Bayes score is : 0.678143712575				
NEGATIVE	0.51	0.46	0.49	228
NEUTRAL	0.74	0.83	0.79	614
POSITIVE	0.76	0.67	0.71	494
Avg/total	0.71	0.71	0.71	1336
Decision Tree score is : 0.710329341317				

E. Prediction Model

The final step in our study is to build a model which predicts where the concerned player is most likely to be transferred to in the future. As a result of the analysis done in the previous module, we have selected SVM (Support Vector Machine) algorithm in order to implement the same as it has the highest accuracy.

Support Vector Machine:-

Support Vector Machines are based on the structural risk minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis „h‟ for which we can guarantee the lowest true error. The true error of „h‟ is the probability that „h‟ will make and error on an unseen and randomly selected test example. An upper bound is used to connect the true error of hypothesis „h‟ with the error of „h‟ on the training set and the complexity of „H‟ (measured by VC dimension), the hypothesis space containing „h‟. SVMs find the hypothesis „h‟ which minimizes this bound on the true error by effectively and efficiently controlling the VC-dimension of „H‟. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. With their ability to generalize well in high dimensional feature spaces, SVMs eliminate the need for feature selection, making the application of text classification considerably easier. Furthermore, SVMs do not require any parameter tuning, since they can find good parameter settings automatically.

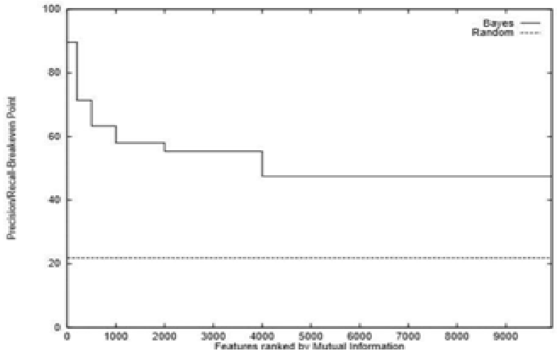

Fig. 1: SVM Learning

We implement SVM on our dataset by loading the previously created CSV file, which contains the player names and the club names, and dividing it into training and test data. We fit our model by giving the training and test data as inputs to the algorithm along with the required labels. Hence, by training the data with the player names and the club names, we can try and predict the club to where any given player is most likely to move to.

4. Implementation

A. Data

The data used for this project involved approximately 15,00,000 raw user tweets collected from the twitter database during the time period of January – April 2017. There were 3 datasets for storing twitter data- one for raw data and the other two for filtered data. The list of clubs used for prediction purposes involved major football clubs from England, Spain, Italy, Germany and France. The list of players involved the players most talked about or mentioned in the users‟ tweets.

The data from the raw tweets was segregated as follows:-

1. The raw tweets were fed into the first parse and filter step in which the dictionaries for players and clubs were used to narrow down the tweets to about 50,000 approx.
2. Further, a second filtering step was used in which all the irrelevant tweets were removed which pertained to a completely different topic. After this step the number of relevant tweets was about 10,000 approx.
3. Further, the analysis module was implemented which narrowed down the tweets to 1336.
4. After the sentiment analysis, only the tweets pertaining to player transfers were retained which amounted to approximately 280.

B. Popularity Mapping

The results obtained from the popularity mapping of football trends on twitter are indicated by the following maps.

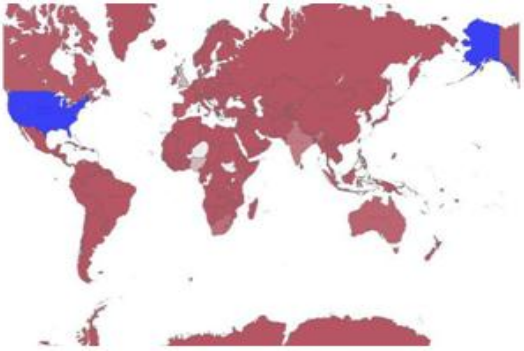

Fig. 2: Density Map for popularity of football

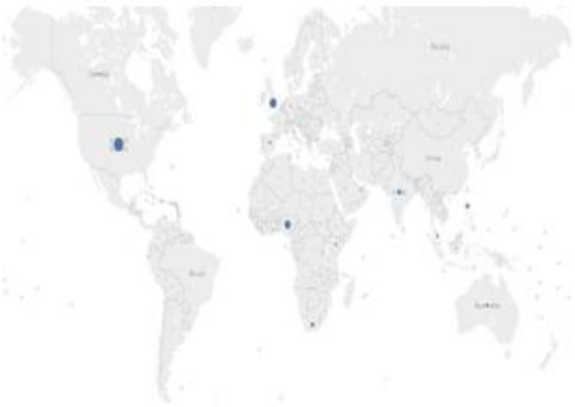

Fig. 3: Dot Map for popularity of football

The maps indicate that the USA has the highest density of football related tweets with 34%, followed by Great Britain and then Canada.

C. Transfer Predictions

The main aim of this module was to predict the player transfers from one club to another with some level of accuracy. The training set and the test set was constructed by feeding the tweets CSV file by means of a package called „sklearn‟. „Sklearn‟ is a package in python which has functionalities such as Count Vectorizer and Tf-idf transformer which help in converting the data in the CSV file to numerical form which is then fed into the SVM algorithm. This model is used to predict the clubs to which the players in the test data set are most likely to be transferred to. A classification report can then be generated with parameters such as Precision, Recall, f1-score and Support. The prediction process involved the use of 2 main dictionaries: list of players and the list of clubs. SVM algorithm was applied on the test and training data sets which were derived from the CSV file containing the players and clubs. The entities involved in this study are some of the most high profile players and clubs from around the world. For e.g. Antoine Griezmann, Arsene Wenger, Mario Mandzukic etc. are the ones which get the most hits on Twitter. Similarly, Manchester United, Barcelona, Real Madrid etc. are the clubs which are most talked about on twitter. The following data indicates the output for the player transfers which are most likely to occur:-

Table 3: Prediction Output

Player Name	Club Name
Antoine Griezmann	Manchester United
Marco Veratti	Barcelona
Massimilano Allegri	Arsenal
Leonardo Bonucci	Manchester City
Mario Mandzukic	West Ham
Zlatan Ibrahimovic	LA Galaxy

Analysis of Football data on Twitter for popularity mapping and transfer predictions, The classification report which was generated after applying SVM on the data yielded an accuracy of 65%, which is pretty reasonable considering the fact that the study was based totally on user tweets and not verified transfer rumours from experts. The following table demonstrates the metrics yielded from the above implementation of SVM.

Table 4: Prediction Output Metrics

Precision	Recall	F-Score
0.57	0.70	0.71
SVM Accuracy is: 0.65		

5. Conclusion

In this paper, we managed to understand more about the trends in football by using user tweets as the basis. We plotted a world map indicating the popularity trends of football around the world by means of twitter filters and Tableau. Further, we predicted the player transfers which could occur in future transfer windows by means of SVM classification model. Hence, we can say that the output from our study helps in understanding more about the game of football and the extent to which the fans play a role in supporting and influencing the decisions made in their favourite clubs.

6. Future Enhancements

The social media analysis on football trends can be expanded further in many ways in order to increase the accuracy. The accuracy tends to increase as the amount of data increases. Hence, with larger amounts of data, the transfer predictions can be done even more effectively. The parse and filter module can be further optimized by implementing the Latent Dirichlet Algorithm (LDA) in order to increase relevancy of tweets. Further, the tweets could be found to be more relevant during the transfer window period (usually June-August). The study can be improved by incorporating bi-grams and N-grams while doing text classification.

References

[1] Min, Byungho, et al. "A compound framework for sports results prediction: A football case study." Knowledge-Based Systems 21.7 (2008): 551-562.

[2] Leung, Carson K., and Kyle W. Joseph. "Sports data mining: predicting results for the college football games." Procedia Computer Science 35 (2014): 710-719.

[3] Brooks, Joel, Matthew Kerr, and John Guttag. "Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights." Proceedings of the 22nd ACM SIGKDD

[4] International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[5] Doman, Keisuke, et al. "Event detection based on Twitter enthusiasm degree for generating a sports highlight video." Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.

[6] Löchtefeld, Markus, Christian Jäckel, and Antonio Krüger. "TwitSoccer: knowledge-based crowd-sourcing of live soccer events." Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia. ACM, 2015.

[7] Van Haaren, Jan, et al. "Analyzing volleyball match data from the 2014 World Championships using machine learning techniques." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[8] Vaz de Melo, Pedro OS, et al. "Forecasting in the NBA and other team sports: Network effects in action." ACM Transactions on Knowledge Discovery from Data (TKDD) 6.3 (2012): 13.

[9] “SupportVectorMachines”,http://scikit-learn.org/stable/modules/svm.html