Soccer Match Analysis

REDDIT THREADS

BY

RICHARD GAO - 260729805
BOGDAN TANASIE - 260747949
ANUKRITI YADAV - 260986353
SIDDHARTH SINGHAL - 260986354

INSY 669- 075 TEXT ANALYTICS

PROBLEM STATEMENT

1. Can we predict the outcome of Pre-Match threads?
2. What are the popular sentiments among the fans of a popular team before a soccer match?
3. What are the popular topics discussed by the fans of the top soccer teams around the globe?
4. Comparing teams from different leagues and know how are they perceived?
5. Can we calculate these things on the fly and display it on a dashboard?

DATA AND ITS SOURCE

The data has been scrapped from one of the most popular social, web content, and rating aggregator website Reddit. The data includes the forum discussion for the game ‘Soccer’. Since our analysis focuses on predicting the outcome of the match using text mining techniques, the data includes the pre-game threads only.

Our dataset consists of 98 unique pre-match threads, having 18327 comments. A user can post multiple comments on a thread. The project currently focuses on comments in the English language. The titles of the match thread have the respective team names with a little description including kick-off time, venue details, and possible line-ups.

DATA PRE-PROCESSING

The data has been pre-processed by defining functions for each of the following techniques in the python notebook:

Replacement

As the data is from a discussion forum, people tend to have multiple ways to represent different teams and to express their opinions. For Example, Manchester United has several nicknames including- The Red Devils, United, ManU. Thus, the team names and the emotions have been replaced in the comments and the title columns of the dataset using predefined data dictionaries in the CSV format.

Tokenization

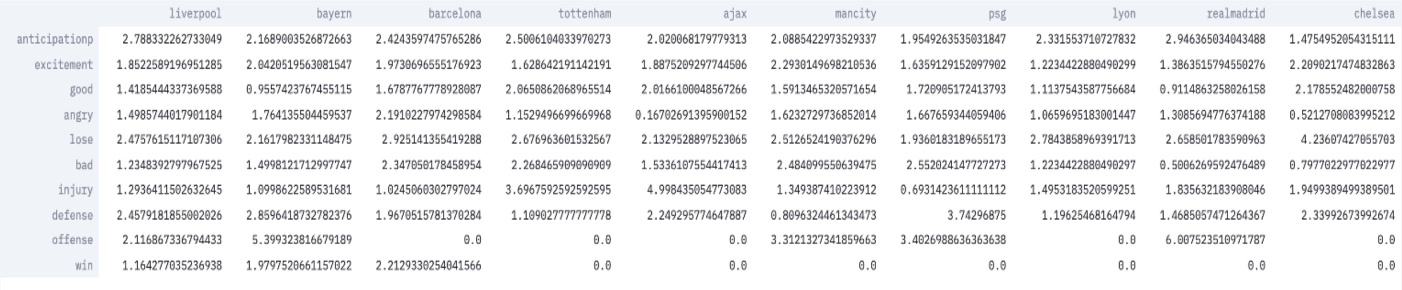
The main part of the data is the comments and the thread title. These columns are textual and unclean. Thus, to clean the data, we first removed every punctuation and special character (if any). We also removed the stop words (e.g.: a, an, the, have, etc.) as these do not add much meaning to a sentence. Then, tokenize the sentences into a list of words. After tokenization, we also created the data time columns i.e. pcreated_date, ccreated_date to include only the comments posted before the match kickoff time.

Lemmatization

As the comments are human written and every person can use different forms of a word, such as performance, performs, and performed. Additionally, there can be derivationally related words with the same meaning. Thus, we use the lemmatization technique to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

EVALUATION METRICS

Lift Values



	liverpool	bayern	barcelona	tottenham	ajax	mancity	psg	lyon	realmadrid	chelsea
anticipationp	2.788332262733049	2.1689003526872663	2.4243597475765286	2.5006104033970273	2.020068179779313	2.0885422973529337	1.9549263535031847	2.331553710727832	2.946365034043488	1.4754952054315111
excitement	1.8522589196951285	2.0420519563081547	1.9730696555176923	1.628642191142191	1.8875209297744506	2.2930149698210536	1.6359129152097902	1.2234422880490299	1.3863515794550276	2.2090217474832863
good	1.4185444337369588	0.9557423767455115	1.6787767778928087	2.0650862068965514	2.0166100048567266	1.5913465320571654	1.720905172413793	1.1137543587756684	0.9114863258026158	2.178552482000758
angry	1.4985744017901184	1.764135504459537	2.1910227974298584	1.1529496699669968	0.16702691395900152	1.6232729736852014	1.667659344059406	1.0659695183001447	1.3085694776374188	0.5212708083995212
lose	2.4757615117107306	2.1617982331148475	2.925141355419288	2.676963601532567	2.1329528897523065	2.5126524190376296	1.9360183189655173	2.7843858969391713	2.658501783590963	4.23607427055703
bad	1.2348392797967525	1.4998121712997747	2.347050178458954	2.268465909090909	1.5336107554417413	2.484099550639475	2.552024147727273	1.2234422880490297	0.5006269592476489	0.7977022977022977
injury	1.2936411502632645	1.0998622589531681	1.0245060302797024	3.6967592592592595	4.998435054773083	1.349387410223912	0.6931423611111112	1.4953183520599251	1.835632183908046	1.9499389499389501
defense	2.4579181855002026	2.8596418732782376	1.9670515781370284	1.1090277777777778	2.249295774647887	0.8096324461343473	3.74296875	1.19625468164794	1.4685057471264367	2.33992673992674
offense	2.116867336794433	5.399323816679189	0.0	0.0	0.0	3.3121327341859663	3.4026988636363638	0.0	6.007523510971787	0.0
win	1.164277035236938	1.9797520661157022	2.2129330254041566	0.0	0.0	0.0	0.0	0.0	0.0	0.0

We related certain words with anticipation, excitement, and anger. In addition, we looked at keywords such as injury, defense, offense etc… This is displayed on the dashboard to get a sense of what is being talked about for team. The user can also input their own keywords to calculate lift score on the fly.

Insights (From Lift Scores):

- People **anticipate** Liverpool matches the most.

- People seem to relate Chelsea and Man United to **losing**.

- People talk a lot about Man United's and PSG's **defense**.

- People seem to be least **angry** at Ajax.

- People mention **good** alongside Ajax.

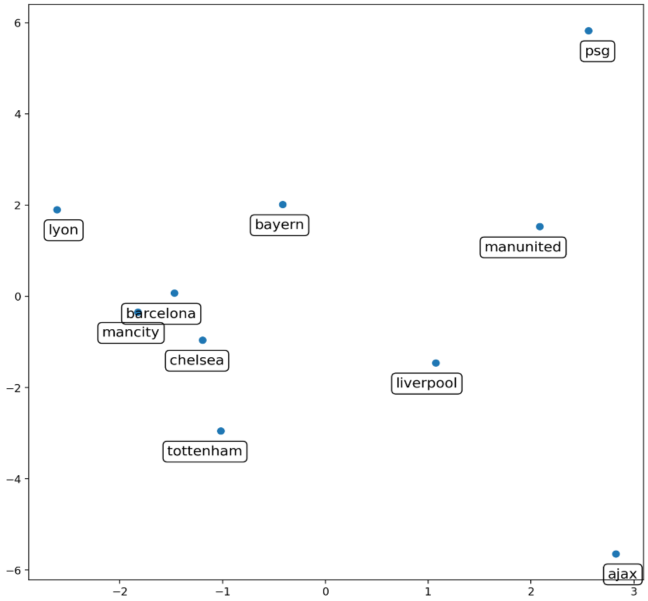
- People are talking about Bayern's **offense**.

- **Injuries** are talked about with Tottenham, Ajax and Porto.

- Barcelona, Man City, and PSG seem attract the word **bad**.

- **Excitement** is highest for Juventus

MDS Plot



Inversed the lift values and calculated the Euclidean distance to create a plot of teams. According the MDS plot for teams greater than 5 and less than 20, it can be observed that the teams from famous leagues like English Premier League (EPL), La Liga-Spanish League, Bundesliga-German football league are close to each other that represents they are top teams in their respective leagues.

Sentiment Analysis

Sentiment Analysis, being the most common text classification tool to classify a message/comment and tell the underlying sentiment is used to calculate the sentiment score based on the comments made by fans on the Reddit forum. We used Sentiment Intensity Analyzer to calculate the sentiments of the comments and the sentiment scores of a team for a thread by grouping them.

Topic Modeling

We first used word cloud to visually identify the most words overall as well as per matchup. (teams vs team) We then used the Gensim Module LDA analysis using both Bag of Words and TF-IDF methods to calculate words frequency. The main goal was to uncover keywords that are used across the different posts as well as for each matchup. For the latter, we believe that it would give us a more granular insight into fans perceptions of different teams and how they face each other. We were able to incorporate these methods into our dashboard to let the potential user explore these methods on their own by modifying parameters such as the matchup. The topic that we uncover are generic (topic 1, topic 2, etc.) and one major improvement would be to use the most important words within a topic to be able to specifically define what the topics are given the context.

INSIGHTS

A sentiment score has been calculated for each comment in a match thread that ranges from -1 to 1 (1 being Positive, -1 negative, and 0 neutral). Similarly, a sentiment score of a match thread has been calculated by grouping the comments using a unique id (pid) for each match thread. From the sentiment score of the match threads, it can be observed that for a one-sided match where one team has way more chances to win from their opponents, the sentiment score is mostly ranging from 0.2 to 0.4. However, if a one-sided match ends with an unexpected result, the sentiment score is negative. That means the team with higher chances of winning underperformed and the fans have posted negative comments.

The Matchup Analysis in the dashboard gives lift and sentiment scores of the respective teams selected in the dropdown menu; Along with a word cloud a prediction for the match is also displayed which is based on the matchup sentiment score. Team with a higher positive sentiment score is predicted to be the winner of the match.

From the current data, it can be observed that the accuracy of the prediction is very low and one of the reasons behind this can be - A team with high social following (Facebook + Instagram + Twitter)[1] will have greater number of comments with positive sentiments towards their team. This creates a bias for a team with high fan following over a less popular but better performing team resulting in less accurate results from the sentiment matchup scores.

FUTURE WORK

Currently, we have considered just the pre-match threads from Reddit with just 98 unique threads i.ie 98 unique matches. Thus, the data may have some biases towards the most popular soccer teams. This problem can be solved by using more historical data, data from other forums, social media posts, blogs and incorporating a variety of teams and leagues data in the model.

Since sentiment and lift values were considered to predict the outcomes of the matches. In our opinion, we feel more techniques and contextual data such as team rank can be used to improve the prediction. Also, the analysis was focused on team-level insights i.e. sentiment scores and lift values were calculated for different teams. It would be interesting to perform a similar analysis on players and managers to deep-dive and gather insights. In addition, being able to figure out which team a comment is describing and perhaps adding the lift score of certain keywords may improve our predictions.

Lastly, to improve the processing time and reduce the manual work, we can explore a few APIs related to soccer to get the match results and other data to analyze and automate the process.

[1] Social Following data- https://withafunfilter.com/the-top-15-biggest-and-most-supported-football-teams-in-the-world/