SENTIMENT ANALYSIS

SUBMITTED BY:

ANSHU PATHAK (670161686)
PRASHANT CHAUDHARY (678226947)
RAJA AMLAN

Purpose

Analyze data from Spanish soccer league popularly referred to as La Liga, to identify the popularity of the top La Liga teams and the sentiments of the fans, using Twitter Hashtags and tweets related to the league and the respective teams.

Background

The Primera Division which is often referred to as La Liga, outside Spain, is the one of the most competitive football leagues in the world. Fans all over the world actively keep tweeting during the course of the league. These tweets may be negative, positive or neutral depending upon the mood of the fans. To analyze the sentiments of the fans, we will use the hashtags and tweets that are relevant to our objective.

Motivation

Each one of us in the group are passionate soccer fans and follow all the major soccer leagues in the world religiously. Ergo, it was an easy choice for us to choose this topic. Further, the project gives us an opportunity to apply big data techniques and concepts that we have imbibed throughout our coursework, thus helping us gain a practical knowledge of the same.

Project Objective

The objective is to find the following:

- Most Popular La Liga Teams in 2018–19 season sorted by Tweet Volume
- Top La Liga Teams in 2018–19 season sorted by Sentiment Score
- Emotional Quotient of Tweets towards the top 5 clubs
- Develop a model to predict the Sentiment of tweets by the users

Tools Used

Python: Tweepy for extracting tweets, TextBlob for sentiment analysis and scores.

Tableau: Visualization

R-Studio: Tweet extraction, NRC Word-Emotion Association Lexicon for analyzing sentiments

Methodology

Data Collection:

We have used Twitter4J API to fetch the Tweets from Twitter based on various hash tags.

1. First, we have started this setup by generating AccessTokens and ConsumerKeys from the Twitter account.

2. Then, created a new Python project with Twitter4j API libraries and used the features of the API to retrieve the tweets from various hash tags.
3. We then identified the required hash tags of the league and the hash tags of the 5 following teams in the tournament.
    - Real Madrid
    - Barcelona
    - Atleltico Madrid
    - Sevilla
    - Valencia

Implementation with Python

In our python script, we analyzed the sentiment of the user. The sentiments were split into the following categories:

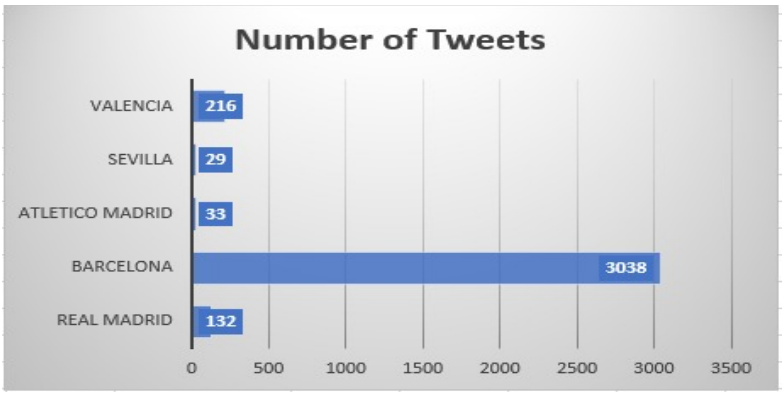
1. Positive
2. Negative
3. Neutral

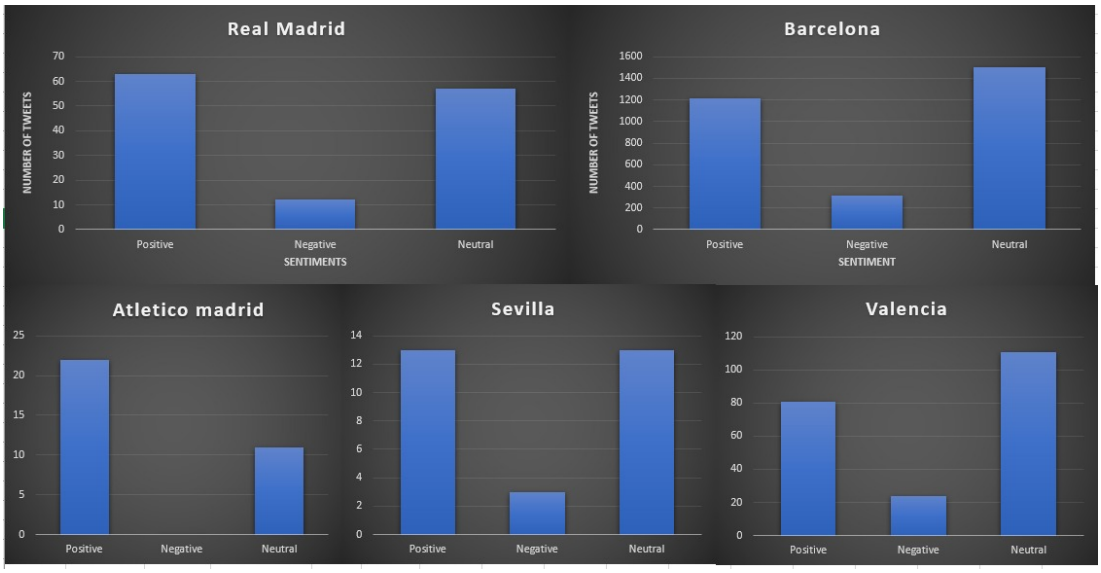
The above analysis was carried out using the TextBlob Library in python.
Further, we developed models which would predict the sentiment of the user based on their tweets.

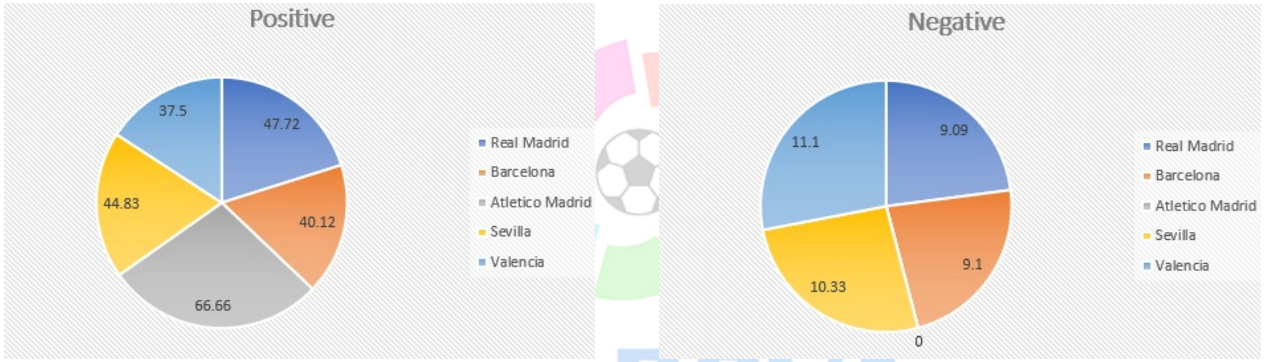
Process

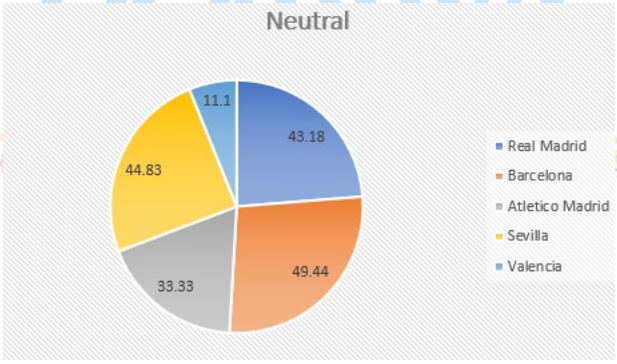
1. Extracted 3500 tweets and preprocessed the data,
2. Preprocessing included, converting all text to lowercase, removing blank space, username, links, tabs and special characters.
3. Mapped the processed data i.e. cleaned tweets to the clubs which they relate to.
4. Utilized the TextBlob library of python, to calculate the sentiment score of the tweets for each club. Textblob comes with the basic features of natural-language processing essentials; we used this for the polarity and subjectivity calculation of tweets
5. Finally, we applied random forest and logistic regression algorithm by splitting the data into training and testing set [80:20], to predict the sentiment of the tweets based on the words used in the tweets.

Below are the snapshots of our result:









Major Inferences:

1. Tweet count for Barcelona is way higher than the others because of their semi-final Leg 2 match in the UEFA Champions League.
2. Valencia has an upcoming semi-final clash against Arsenal in the Europa League, so a lot of fans tweeted in support for the club, which is why the percentage of positive tweets for Valencia is the highest as compared to other clubs.
3. Real Madrid's season did not go as expected; hence, we see slightly negative emotions coming from the fans. Its, positive tweet is likely because of the confidence in their newly appointed manager. Atletico Madrid did pretty well throughout the year, and so we did not see any negative tweets for the club.
4. All the fans of the respective clubs have almost same percentage of neutral sentiment

Implementation with R

Objective: Analyzed the sentiments of the users behind the tweets based on the NRC Word-Emotion Association Lexicon. The sentiments are divided into following categories:

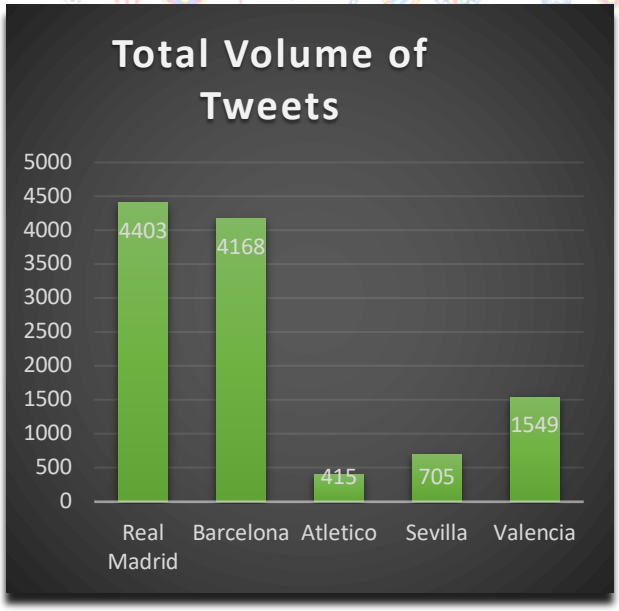
1. Positive
2. Trust
3. Anticipation
4. Negative
5. Joy
6. Surprise
7. Sadness
8. Fear
9. Anger
10. Disgust

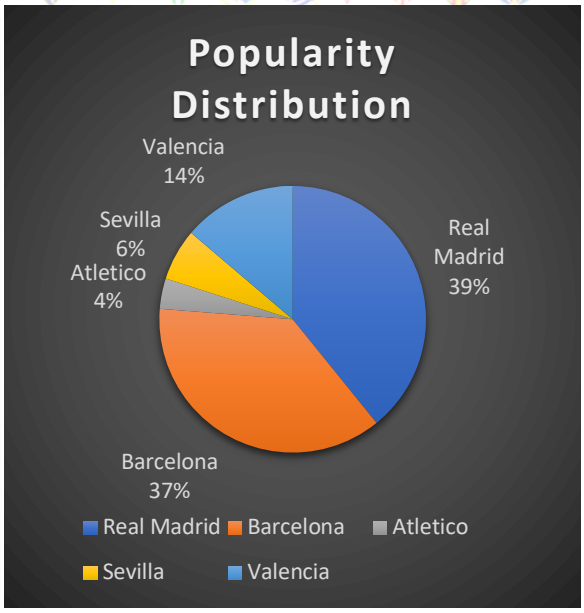
Python v R: The results and visualization would be different in R as compared to Python because the tweets were extracted on different days for R and Python.

Process

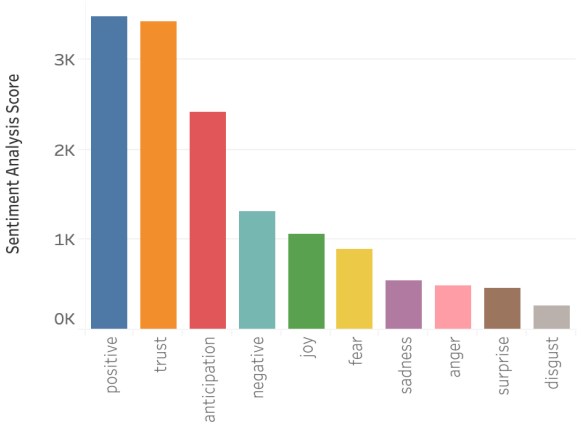
1. Tried extracting 25k tweets, but due to the rate limit of twitterAPI we could only get around 11k tweets.
2. Applied data cleaning techniques such as converting all text to lower case, removal of blank space, hyperlinks, punctuations and stop words
3. Analyzed the sentiments of the users behind the tweets based on the NRC Word-Emotion Association Lexicon

Below are the visualizations of our analysis:

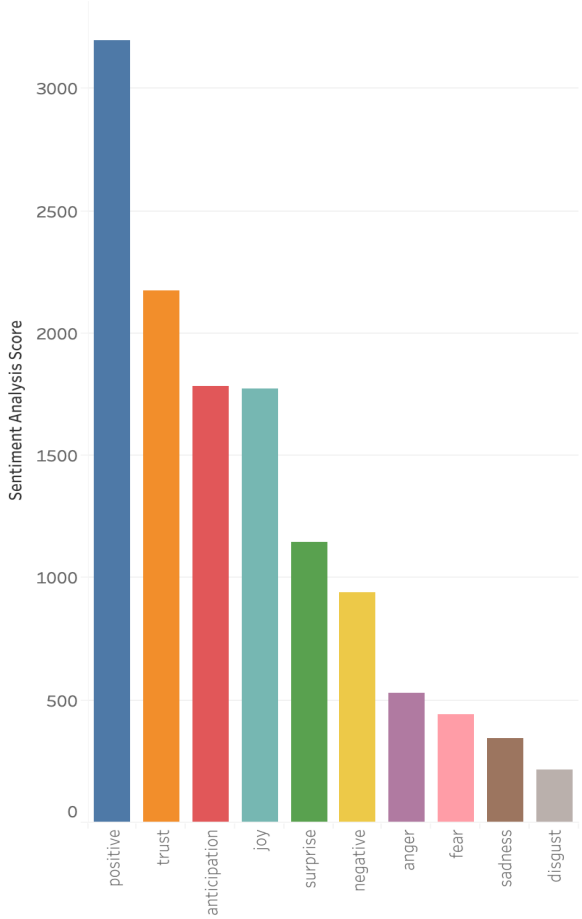




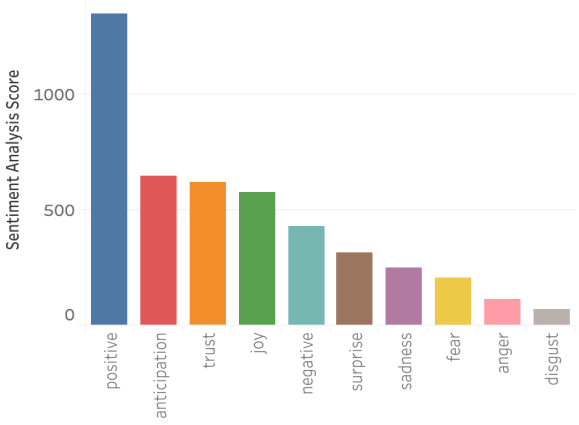
Real Madrid

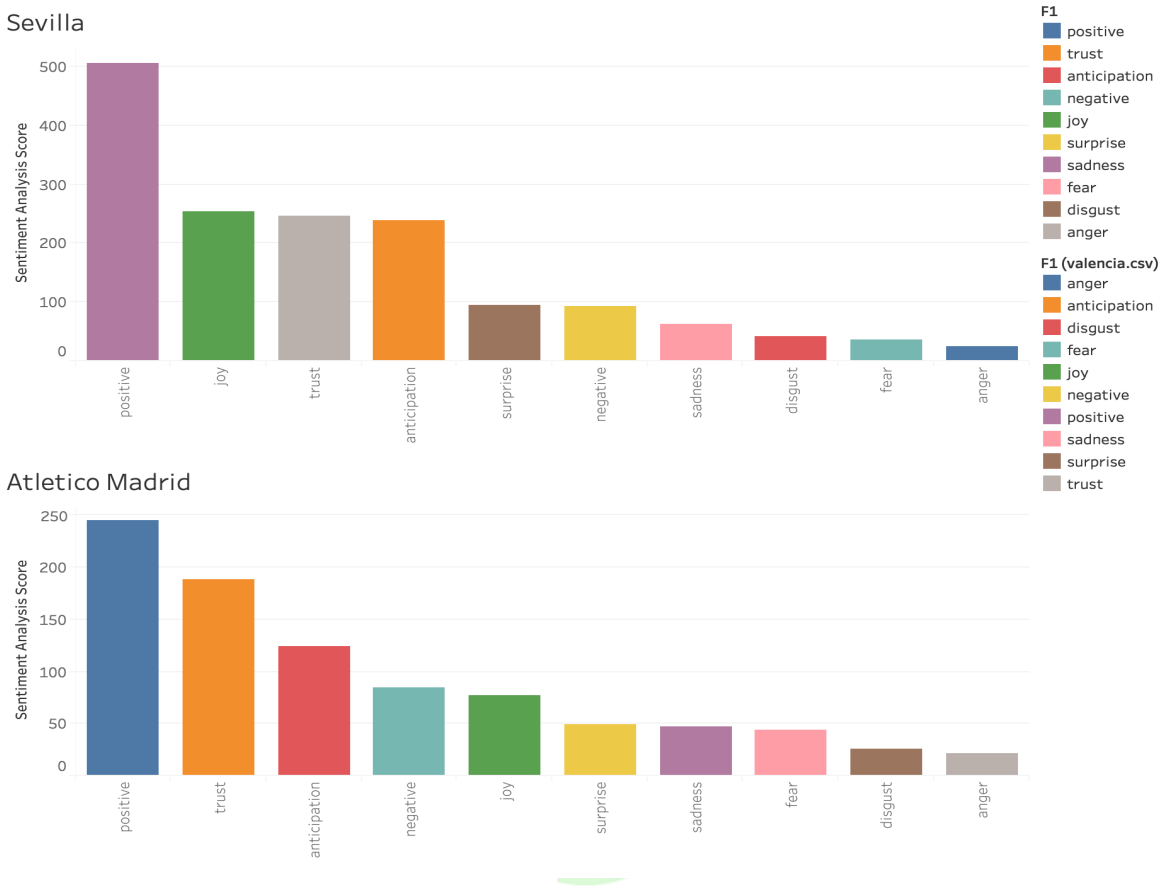


Barcelona



Valencia





Major Inferences:

1. We observe that the tweet count for Real Madrid were highest followed closely by Barcelona. Valencia, Sevilla and Atletico had comparatively a smaller number of tweets.
2. Looking at the world map, we see that most of the tweets originated from Europe as expected, followed closely by India and countries in South America where the viewership of the league is of considerable amount.
3. By analyzing the emotional quotient of the fans, we see that Real Madrid fans tweeted very positively which is likely due to the positive news regarding one of the legendary players doing well after a suspected heart-attack. Additionally, this may have lead to the high number of anticipatory tweets too. Further, re-appointment of Zidane as its new manager could have lead to the high number of trust tweets.
4. Further, Barcelona had considerable amount of positive tweets which was due to their success in the Leg 1 of the semi-final match in the UEFA Champions League.
5. Valencia also has a high number of positive tweets due to its recent success in the Europa League.

Conclusion

Based on the above analysis, we can conclude that the sentiment level for the top 5 teams vary depending upon the circumstances. Usually, the team with upcoming fixtures in the nearest future, tends to be more talked about than the others, provided important and noteworthy events don't take place with respect to some other club. The above analysis can be put to an effective use for advertising and other business

purposes, as one can set the target audience depending upon the sentiments of the fans, who in fact are the ultimate viewers. Further, a sentiment analysis of leagues can be done country-wise, for a more targeted marketing approach.

References:

https://www.kaggle.com/priyaananthram/sentiment-analysis-of-tweets

https://medium.com/@GireeshS/sentiment-analysis-of-tweets-during-ipl-2018-finals-78f6e940f3d0

https://github.com/byam/predictEPL

https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/