**Capstone Project – FIFA World Cup 2018 Twitter Analysis**

**Miles Rousseau**

**Submission date: 11/7/22**

Problem Statement:

FIFA should aim to constantly improve the fan experience of the FIFA World Cup. Fans throughout the world express their opinions of aspects of the World Cup through Twitter and thus there is a large amount of data that can be analyzed to gain insights into the opinions of fans and be used to paint a picture of the overall fan experience. This data analysis should be used to help FIFA improve the organization and promotion of the World Cup. I have conducted an in-depth analysis of a data set of tweets from the last World Cup in 2018 in order to help FIFA improve on their core responsibilities and create suggestions for the upcoming World Cup in Qatar in November 2022.

Introduction:

The FIFA World Cup is one of the most watched sporting events in the world, and millions of fans worldwide chime in on Twitter to voice their opinions on the proceedings. With the next world cup approaching this November 2022 in Qatar, I will be looking into what fans on Twitter had to say about the last installment of the World Cup in 2018. The goal of this analysis is to gain insights about the likes and dislikes fans expressed on Twitter to create suggestions that FIFA can use to further improve the fan experience. I used a variety of natural language techniques such as sentiment analysis and topic modelling and key data exploration/visualization techniques along with my knowledge of the subject matter to investigate the problem statement.

Dataset:

I found and downloaded my FIFA World Cup Tweet dataset from this Kaggle upload: <https://www.kaggle.com/datasets/rgupta09/world-cup-2018-tweets>. The data is a collection of 530K tweets about the competition starting from the Round of 16 until the Final, 6/29/22 until 7/15/22. The dataset uploader used Tweepy API to collect tweets using various key hashtags to filter for world cup specific tweets. This dataset is a sample of all tweets made during this period and does not have every single tweet made about the world cup during this time period. The dataset had been pre-processed to only include English language tweets and had been cleaned of all non-language elements, such as website names, special characters, retweets, user mentions and hashtags.

The data was downloaded as a .csv file. The original file contained 16 columns with key columns being a string of the tweet itself, the datetime of tweet, the location of tweet as a string, the hashtags and mentions as strings, and the number of followers, friends, likes, characters, and retweets as ints. The dataset had 530,000 rows with null values in only a select few columns, most importantly the cleaned tweet column had 551 missing tweets and these rows were dropped. Therefore, the initial shape of the schema was 16 columns and 529499 rows.

Data Cleaning and Preprocessing:

The first order of businesses was dropping the rows containing missing tweets as previously discussed. After that, basic preprocessing was performed which included changing the date data type to a pandas datetime data type, removing unnecessary columns which added no useful information like the language column which contained the same value (English) for each row. The ID, usermentionID and Name columns were also dropped because they added no useful information.

Table

Description automatically generatedAfter basic preprocessing, the Place column containing data on where the tweet was uploaded from needed a lot of cleaning. Tweet location was organized by city, state, country, and a combination of all three in some cases. To simplify the column, I extracted the country where the tweet was made in the applicable cases and assigned the value to ‘Other’ where it was impossible to determine the origin. In the end, 69 countries were identified with 239,769 tweets and the remaining 289,680 tweets were classified as coming from ‘Other’. I also filled in the null values in the hashtags and user mentions columns to None, as this represented tweets with no hashtags or no mentions.

Initial DataFrame

Table

Description automatically generated with medium confidence

Processed DataFrame

Insights, Modelling, and Results:

After cleaning my data, I was ready to assign sentiment to each tweet to get a better understanding of how fans felt about various topics. For the sentiment analysis, I used Vader sentiment analyzer from sklearn to classify my tweets into negative, neutral, and positive.

Icon

Description automatically generatedMost tweets in this dataset were classified as positive which is great news to FIFA as this indicates that the World Cup in 2018 was received well based on this initial look. However, there is much room for improvement given the nearly 80,000 negatively classified tweets. 54.7% of all tweets in the dataset were considered to be positive, 32.1% were neutral and 13.1% were negative.

After classifying tweet sentiment, I investigated how topics, people and hashtags were viewed. To do this efficiently, I vectorized the tweet, hashtags and mentions columns and then visualized how various words interacted with the sentiment of the tweet. Using a tokenizer I created, I pulled out key words from tweets and mapped how these words were associated with negative, positive, and neutral tweets. Additionally, I looked at what sentiment was correlated with various hashtags and user tags. Some of the most interesting insights in the initial EDA came from looking at the semi-finalist teams of the competition. England, Croatia, France, and Belgium were the last four teams in the tournament, and I looked at the general sentiment of the tweets with the hashtags of these countries in the dataset. Both England and Belgium lost their semi-finals, and this is clearly reflected by the percentage of tweets that are classified as negative when using the country’s hashtag. England specifically had a significant number of negative tweets which can be explained by a large portion of the tweets in the dataset coming from England as the dataset only contained English language tweets.

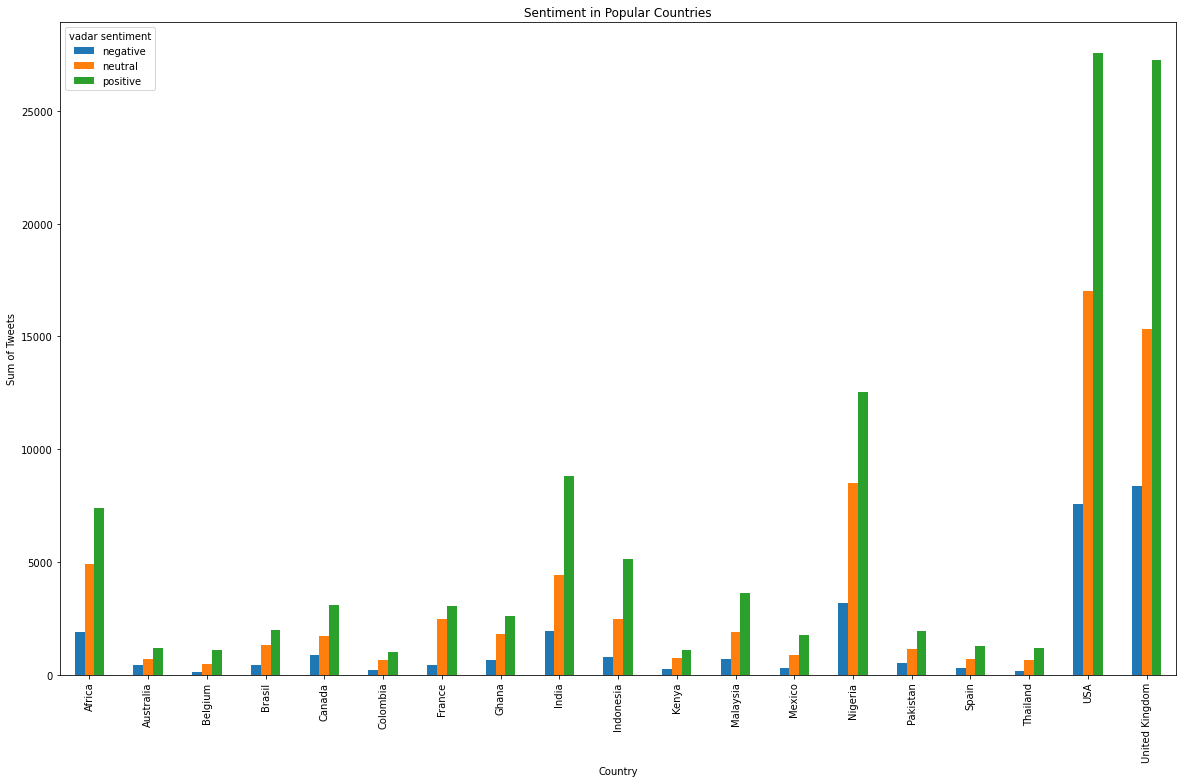
**Sentiment of Tweets using Semi-Finalist Hashtags**

Icon

Description automatically generated



A positive trend in the EDA phase that was noticed, was that the proportion of tweets that were positive from each country that the tweet was posted from was overwhelmingly positive, showing that fans from all over the world similarly enjoyed the World Cup. A look into the sentiment from the top 19 tweet origin countries can be seen below.



Findings and Conclusions:

From the data analysis I gained insight into 5 data driven suggestions that FIFA can use to help organize and promote the upcoming FIFA World Cup in Qatar.

Suggestions:

* The first installment of using VAR was positively received and should continue on in FIFA organized tournaments.
  + 80% of tweets about VAR had positive or neutral sentiments, when compared to the overall topic of refereeing which had 64%, this a significant statistic.
* Kylian Mbappe can be used as the new poster boy for the upcoming world cup.
  + Mbappe was the most mentioned player in the entire dataset, and had 15,000 positive tweets in which he was mentioned, significantly greater than the former poster boy Messi who only had 7,000 positive tweets.
* For the English-speaking markets, England is the most important market for promotions as it has the largest worldwide support.
  + During the World Cup Knockout stages, the result of the England team swayed the overall sentiment of the tweets based on whether they won or lost. On the day they won the quarter final, sentiment compound score for all tweets rose by 0.1 and when they lost in the semi final, sentiment score dropped by 0.15 compared to the previous days.
* Nigeria is an emerging market that interacts constantly on twitter and should be looked into for marketing promotions.
  + Though Nigeria isn’t a ‘traditional’ footballing country, it had the third most tweets of any country and the third most retweets.
* FIFA should aim to engage with the community during high impact days.
  + Most World Cup related tweets are directly correlated to when games are being played. On every game day, twitter engagement spikes and this is a great time for FIFA to engage with fans to overall better their experience and excite engagement with the brand.

In conclusion, there is a lot of useful data regarding brands and events on Twitter. This data comes in the form of text but using the right data science techniques can be transformed into useful information. My twitter analysis on the World Cup 2018 showed me how much information can be gained from looking at tweets. From this basic analysis, FIFA can create data driven actions to improve fan experience, improve organization of tournaments and optimize promotional strategies. However, I only sampled a few topics that I found interesting and there is a multitude of different topics that can be explored to further paint the picture of what fans thought about the FIFA World Cup 2022. My intended next steps for this project are to expand the range of topics explored and potentially look into live tracking of sentiment for the upcoming FIFA World Cup in Qatar.