Correlation of Twitter Sentiment Analysis and 2022 FIFA World Cup Performance

Graz University of Technology
Master of Computational Social Systems
Foundations of Computational Social Systems
Winter Term 2022/23

Tse Yan Lui
Farzana Akter

Motivation

The inspiration for this project arises from the increasing relevance of social media in our lives and the role it plays in influencing public opinion. Twitter is a real-time venue for individuals to express their views and emotions, and it has become an appropriate source of information for gauging public mood. The 2022 FIFA World Cup, as one of the world's largest and most-watched sporting events, provides a unique chance to explore the relationship involving people's emotions and the performance of their favorite teams.

Furthermore, analyzing Twitter user sentiment during the FIFA World Cup can give useful insights into the relationship between public opinion and match outcomes. By completing this study, we aim to get a deeper understanding of the link between people's emotions and the performance of their favorite teams, as well as how this relationship affects game outcomes. This information has enormous significance for sports organizations, sponsors, and even for team-boards, since it may help them get a better knowledge of public sentiment and taking better initiatives.

Consequently, the goal of this project is to intensify our understanding in the field by investigating the relationship between Twitter user sentiment and FIFA World Cup performance, as well as to contribute to a greater understanding of the role of social media in affecting public opinion.

Data Retrieval

The data was collected from FBref.com [1]. It is a website dedicated to recording statistics for football clubs and players all around the world. FBref was built by Sports Reference, the organization behind prominent statistic websites such as Baseball-Reference and Basketball-Reference.

Retrieved data was then analyzed to find any links between the sentiment of the tweets and the performance of the teams, as assessed by the (**Gls** - **GA**)/2 per game computation. Columns **Gls** is stands for Goals Per Game and **GA** stands for Goals Against Per Game.

To retrieve data for this project, we have used the Python package "snscrape" [2], which is a tool for easily scraping data from social media networks like Twitter, Facebook, Instagram, Telegram, Reddit, and so on. It scrapes information such as user profiles, hashtags, or searches and delivers the results, likely relevant postings. One of the most important characteristics of "snscrape" is its ability to execute real-time scraping, which allows you to obtain the most recent social media data as it is uploaded.

Data processing

Filter tweets

	Country	Datetime	Tweet Id	Text	Username
0	Argentina	2022-12-17 23:55:02+00:00	1604263934730854406	Ex-Manchester United captain predicts the winn...	sportnewsblogd1
1	Argentina	2022-12-17 23:48:26+00:00	1604262273543176192	@FIFAWorldCup This it's gonna be tight game an...	im_abhay4u
2	Argentina	2022-12-17 23:48:06+00:00	1604262191812722688	@FIFAWorldCup While everyone is waiting for th...	i8u9i
3	Argentina	2022-12-17 23:46:08+00:00	1604261695597207557	@FIFAWorldCup Don't mess it up with your unfai...	Bal_Aissata
4	Argentina	2022-12-17 23:46:06+00:00	1604261686399176704	Mbappe please humble Argentina and messi! Save...	md7y_

TwitterSearchScraper, a model from Snscrape, was used to scrape tweets by searching for matching specific keywords, hashtags, or phrases. The data was collected using a query that includes the hashtags “FIFAWorldCup” and “country” from “since:2022-11-20” to “date”. Where “country” is the country that participle in the World Cup and “date” is the end date of the tweet collected. The “date” is different for each country based on the last game it played.

The TwitterSearchScraper retrieved tweets that matched the provided query and country and added important information to the list, such as tweet id, content, and username. This produced a final DataFrame including all of the tweets along with “Country”, “Datetime”, “Tweet id”, “Text”, and “Username”.

Normalized values & computed additional variables

	Country	Datetime	Tweet Id	Text	Username	Tweets_clean
0	Argentina	2022-12-17 23:55:02+00:00	1604263934730854406	Ex-Manchester United captain predicts the winn...	sportnewsblogd1	united captain the winner of world cup 2022 fi...
1	Argentina	2022-12-17 23:48:26+00:00	1604262273543176192	@FIFAWorldCup This it's gonna be tight game an...	im_abhay4u	this its be tight game and will 🏆ar💙
2	Argentina	2022-12-17 23:48:06+00:00	1604262191812722688	@FIFAWorldCup While everyone is waiting for th...	i8u9i	while everyone is waiting for the world cup fi...
3	Argentina	2022-12-17 23:46:08+00:00	1604261695597207557	@FIFAWorldCup Don't mess it up with your unfai...	Bal_Aissata	dont mess it up with your unfair referee in fa...
4	Argentina	2022-12-17 23:46:06+00:00	1604261686399176704	Mbappe please humble Argentina and messi! Save...	md7y_	please humble and save humanity from an upcomi...

Text normalization, since the tweets could contain many items, including plain text, mentions, hashtags, links, and more. After the removal processing, it might result in empty fields during the cleaning process, as some tweets only contain links or hashtags that confuse a computer model. These blank rows would delete too. And will fill with a new cleaned tweet to ensure all extracted country tweets are in the same amount.

Since there are in total of 37 countries that participated in the World Cup, it is arduous to fit a model that has numerous annotated corpora to analyze all these languages. Therefore, we decided to remove the non-English tweets. We didn’t remove the emoji since the selected model can process emojis. And emojis are widely used in social media to convey emotions, opinions, and reactions, providing a rich source of sentiment information in addition to the text tweet.

Analysis

	Country	Datetime	Tweet Id	Text	Username	Tweets_clean	VADER_class	VADER_score
0	Argentina	2022-12-17 23:55:02+00:00	1604263934730854406	Ex-Manchester United captain predicts the winn...	sportnewsblogd1	united captain the winner of world cup 2022 fi...	Positive	0.7650
1	Argentina	2022-12-17 23:48:26+00:00	1604262273543176192	@FIFAWorldCup This it's gonna be tight game an...	im_abhay4u	this its be tight game and will 🏆🇦🇷💙	Positive	0.6369
2	Argentina	2022-12-17 23:48:06+00:00	1604262191812722688	@FIFAWorldCup While everyone is waiting for th...	i8u9i	while everyone is waiting for the world cup fi...	Positive	0.5994
3	Argentina	2022-12-17 23:46:08+00:00	1604261695597207557	@FIFAWorldCup Don't mess it up with your unfai...	Bal_Aissata	dont mess it up with your unfair referee in fa...	Positive	0.6261
4	Argentina	2022-12-17 23:46:06+00:00	1604261686399176704	Mbappe please humble Argentina and messi! Save...	md7y_	please humble and save humanity from an upcomi...	Neutral	0.0258

VADER sentiment analysis was used to determine the sentiment of the collected tweets. The VADER model analysis tweets by VADER class and VADER score. The VADER class label tweets as a positive, negative, or neutral class. The VADER score represents the intensity of mood from -1 to 1, where -1 indicates a strong negative correlation with emotion and 1 indicates a strong positive correlation with emotion.

	Country	Mean_score	Median_score	Std_score
0	Argentina	0.351650	0.4404	0.418743
1	Australia	0.217420	0.0129	0.471765
2	Belgium	0.193557	0.0387	0.420133
3	Brazil	0.260088	0.2500	0.436248
4	Cameroon	0.271504	0.2023	0.404690

In the context of the World Cup, mean, median, and standard deviation are used to analyze the sentiment scores of tweets from different countries, providing insights into how fans of each country are feeling about their team's performance during the tournament. Using these three scales represents a comprehensive understanding of the sentiment distribution in the data. The mean represents the average sentiment score of the tweets, providing a general idea of the overall sentiment trend in the data. The median is the middle value of the dataset which is effective in analyzing the data if outliers distort the mean. The standard deviation measures the spread of the data and provides information on how consistent the sentiment scores are within the dataset.

	Country	MP	Gls	GA	Performance
0	Argentina	7	2.14	1.14	0.500
1	Australia	4	0.75	1.50	-0.375
2	Belgium	3	0.33	1.50	-0.585
3	Brazil	5	1.60	0.60	0.500
4	Cameroon	3	1.33	1.33	0.000

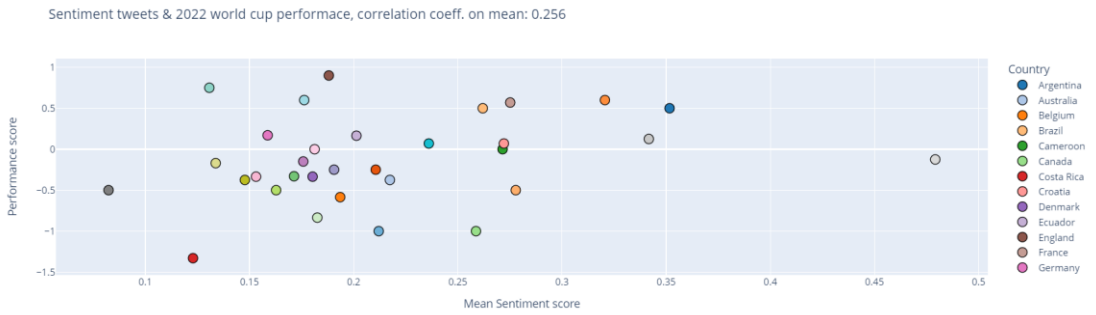
Gls, goals per game, and GA, goals per game, are used to evaluate team performance. The performance measure is calculated by dividing the difference between Gls and GA by 2. The measurement represents the team's overall performance in the World Cup and provides an idea of a team's offensive and defensive strength. This calculation provides a complete picture of the team's performance in the World Cup and evaluates the country's performance against other teams.

	Country	Mean_score	Median_score	Std_score	MP	Gls	GA	Performance
0	Argentina	0.351650	0.44040	0.418743	7	2.14	1.14	0.500
1	Australia	0.217420	0.01290	0.471765	4	0.75	1.50	-0.375
2	Belgium	0.193557	0.03870	0.420133	3	0.33	1.50	-0.585
3	Brazil	0.261997	0.25445	0.435939	5	1.60	0.60	0.500
4	Cameroon	0.271504	0.20230	0.404690	3	1.33	1.33	0.000

Pearson correlation	Correlation coefficient	P-value
Mean & performance	0.25640495928776574	0.15661534183676534
Median & performance	0.34641373053373425	0.05210516279663612
standard deviation & performance	0.1188906575895279	0.5169240864080457

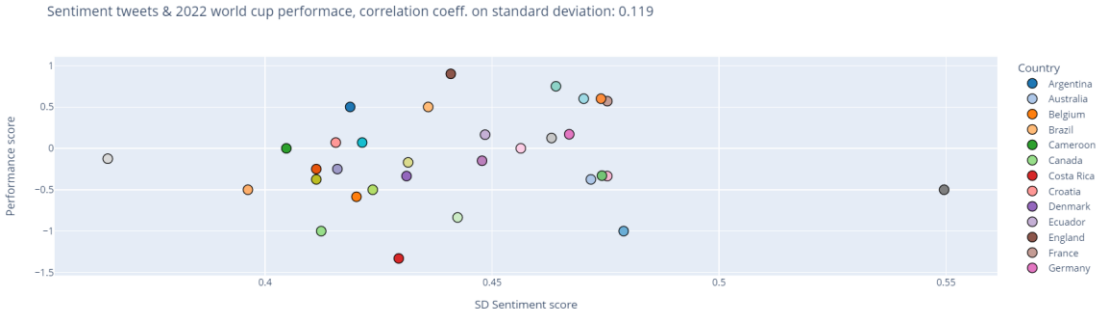
Pearson correlation was performed to analyze the relationship between the mean, median, and standard deviation score of tweets" sentiment and the country's performance. The Pearson correlation coefficient measures the strength and direction of the linear relationship between two variables. The coefficient ranges from -1 to 1, where -1 represents a strong negative correlation, 0 represents no correlation, and 1 represents a strong positive correlation. And the p-value is a measure of the statistical significance of the correlation.

Also, creating the scatter plots for data visualization to visualize the correlation between the variables.



The correlation coefficient of the mean sentiment score and the performance score is 0.256. The mean tweet's sentiment plots of the country are mostly centralized around the range in sentiment scores 0.1 to 0.35. Iran had the lowest mean sentiment score, 0.082, with -0.5 performance score. The United States had the highest mean sentiment score, 0.479, with -0.125 performance score. The correlation coefficient represents there is a weakly positive relationship between sentiment and performance.

The correlation coefficient of the median sentiment score and the performance score is 0.346. The coefficient shows a weak positive relationship between these two variables. However, there are 8 countries, Spain, Uruguay, Mexico, Poland, Saudi Arabia, Korean Republic, Switzerland, and Costa Rica, that are laid on 0 sentiment score. 0 is the lowest median sentiment score. While both Costa Rica and Spain had a 0 median sentiment score, the performance score of Spain was 0.75 while Costa Rica was -1.33. As we removed the non-English tweets for data processing, it might affect the accuracy of the sentiment score, resulting in a 0 median sentiment score for this country. Again, the United States had the highest median sentiment score, 0.523, with -0.125 performance score.



The correlation coefficient of the standard deviation sentiment score and the performance score is 0.119. The coefficient represents a weak positive relationship between these two variables. The standard deviation tweet's sentiment plots of the country are mostly centralized around the range in sentiment scores 0.4 to 0.5. The United States had the lowest standard deviation sentiment score, 0.365, with -0.125 performance score. Iran had the highest standard deviation sentiment score, 0.550, with -0.5 performance score.

Conclusion

In conclusion, Pearson's correlation results between the mean, median and standard deviation of sentiment scores and performance showed a weak positive correlation with values of 0.256, 0.346, and 0.119 respectively. However, since all the p-value are larger than 0.05, suggesting there is a larger than 5% probability that the correlation

between sentiment scores and performance scores is randomness. It might be intriguing to claim that there is a correlation between sentiment and performance. It is important to remember that correlation only measures the strength and direction of a linear relationship between two variables and that there may be other non-linear relationships between mood scores and performance scores that are not explained by correlation coefficient.

The full code and dataset used in this report can be found in the Github repository [3]. The repository serves as an additional resource for stakeholders who want to learn more about the project.

Critique

The performance of a World Cup team is a complex issue and can be influenced by many factors other than public opinion, such as team dynamics, individual player performance, tactics, and luck. These factors may not be directly captured in the sentiment scores, which may limit the ability of the analysis to accurately reflect the relationship between sentiment and performance.

The analysis provides a preliminary insight into the relationship between public sentiment and world cup performance, but more comprehensive studies with larger data sets and more advanced methods are needed to validate the findings. The correlation results could be impacted by factors such as small sample sizes and the selection and representation of tweets. It should note that correlation does not imply causation and that there may be other factors affecting performance scores that are not explained by sentiment scores.

Besides, NLP models used for sentiment analysis may have biases and limitations. The sentiment scores obtained in this analysis only consider sentiment expressed through tweets, and may not accurately reflect the sentiment of the majority of the population. Tweets can be cynical, humorous, or misleading which may not accurately represent the emotion of the user. Therefore, it is crucial to consider of the limitations of the NLP model and the data used to interpret the results of this analysis.

Last but not least, while the weak positive correlation between sentiment scores and performance scores suggests that public opinion may have some influence on a team's performance, it is not possible to draw a definitive conclusion based on this limited analysis. The analysis provides a preliminary insight into the relationship between public sentiment and world cup performance, but more exhaustive studies with larger data sets and more advanced methods are needed to validate the findings.

References

1. FBref. 2022. 2022 FIFA World Cup Stats. https://fbref.com/en/comps/1/stats/World-Cup-Stats

2. JustAnotherArchivist. 2022. Snscrape. https://github.com/JustAnotherArchivist/snscrape

3. Lui-Tse. 2023. FCSS-project. https://github.com/Lui-Tse/FCSS-project