The FIFA World Cup 2018 Supporter Prediction Based on Twitter

Najnin Jahan
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh
njahan222023@mscse.uiu.ac.bd

Md. Hasnat Abdullah
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh
mabdullah223022@mscse.uiu.ac.bd

Tanvir Hossain
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh
thossain222016@mscse.uiu.ac.bd

Abstract— **The FIFA World Cup is an international football tournament that captivates the attention of millions of fans around the globe. In 2018, the tournament took place in Russia, and the fervor and excitement surrounding the event were unparalleled. This focuses on the prediction of supporter behavior during the FIFA World Cup 2018, aiming to understand the factors influencing fan support and identify patterns that could aid in forecasting fan preferences. To achieve this, a multi-faceted approach was employed, including historical team performance, geographical factors, social media sentiment analysis, and previous fan engagement metrics. Machine learning algorithms, deep leaning algorithms and also clustering is applied to analyze and process the collected data to generate insights and predictions. This abstract provides a brief overview of a comprehensive study conducted on predicting supporter behavior in the FIFA World Cup 2018. By employing deep learning techniques and machine learning algorithms. And got the best accuracy from Decision Tree which is 88% and lowest accuracy from LSTM which is 31%.**

Keywords— Prediction, classification, deep learning, football, clustering, preference, team, tournament.

INTRODUCTION

When it comes to major international football tournaments like the FIFA World Cup, the passion and excitement among fans are palpable. While the matches themselves are filled with suspense and unpredictability, another thrilling aspect of these tournaments is the phenomenon of supporter predictions. Fans from all corners of the world gather, both physically and virtually, to offer their insights, forecasts, and speculations about the outcomes of matches, often leading to heated debates and friendly competitions. The FIFA World Cup is an international football competition. People support different teams in the FIFA World Cup by watching the player's skill, the number of trophies they won and the deliberate tricks applied by these teams during the tournament [1]. This is what people post about on social media during the game. By watching these we can predict the fans and team preferences of the game. The FIFA World Cup, the pinnacle of international football tournaments, captivates billions of fans around the globe. With its rich history and a legacy of memorable moments, the World Cup showcases the unparalleled skill and dedication of the world's top footballing nations. However, it is not only the actions on the pitch that make the World Cup a truly remarkable event; the unwavering passion and fervor of the supporters are equally integral to the tournament's magic. Football has an unparalleled ability to bring people together, and the shared experience of watching matches ignites a sense of camaraderie among fans. It is in this spirit that supporter predictions gain significance. Whether it's discussing potential winners, analyzing team performances, or debating the impact of key players, fans engage in a whirlwind of conversations, forming their own hypotheses about the future course of the tournament [2]. This collective anticipation and the accompanying sense of friendly competition on social media like twitter add an extra layer of excitement to the already thrilling spectacle of FIFA tournaments. Supporter predictions are not merely based on blind faith or personal biases. Fans meticulously analyze historical data, team statistics, player performances, and tactical strategies to arrive at their predictions. While some rely on gut feelings and intuition, others adopt a more data-driven approach, utilizing advanced statistical models and machine learning algorithms to assess the probabilities of different outcomes [3]. The diversity of methodologies employed by fans showcases the extent of dedication and creativity within the supporter community.

A. Objectives

1. To analyze the dataset of tweets collected from Twitter during the 2018 FIFA world cup and gain insights into people's social media activities and preferences.
2. To identify the data properties and relationships between the features in the given dataset.
3. To predict which teams are likely to be the most popular in future tournaments.
4. To demonstrate the value of data analytics in extracting valuable insights from large and complex datasets and making data-driven decisions.

B. Contributions

- We are going to predict users team preference. Also, we are the first to find out differences and similarities between fans according to their tweets.
- We built four different classification models like Decision Tree, KNN, Random Forest, Adaboost.
- Balanced the data using SMOTE.
- Worked with both Linear regression and Logistic regression.
- Also built deep learning models like LSTM and RNN.

To get the best performance of proposed models, a large amount of data is required for proper training and testing of the model. Due to the irrelevant data of the dataset, it is essential to balance the data and preprocess which is another main objective of this study besides predicting supporters. In this study, we have used Decision Tree, KNN, Random Forest, RNN, LSTM and Adaboost to predict supporter form tweets. To evaluate the performance of the working models, we calculated accuracy and other performance evaluation metrics: precision, recall and F1-score.

Related work

Several studies have explored the prediction of supporter preferences and behavior during the FIFA World Cup 2018. Here are a few relevant works. This research is inspired by some previous work in this related field and some to our knowledge. Over the past few years, several investigations have been carried out concerning the prediction of sports outcomes through the utilization of various machine learning methods [4]. One notable study by Nichols et al. [5] focuses on developing a sports summarization technique derived from tweets. Another study conducted by Leung et al. [6] concentrates on forecasting game results by leveraging historical data. These research efforts have contributed to the growing body of knowledge in the field of sports result prediction using machine learning techniques. This study [7] focuses on predicting match outcomes during the FIFA World Cup 2018 using machine learning techniques. While it primarily concentrates on game results, it indirectly contributes to understanding supporter behavior by incorporating factors such as team popularity and player performance. This research [8] examines Twitter data to predict the sentiment of FIFA World Cup 2018 supporters. By analyzing the sentiments expressed in tweets related to the World Cup, the study provides insights into supporter preferences and emotional responses towards different teams and matches. The article [9] explores fan sentiment analysis and team performance prediction during the FIFA World Cup 2018. By analyzing fan sentiments from social media platforms, the study sheds light on supporter preferences, expectations, and reactions to various teams' performances. The country's team's sport of choice is soccer, and its impact on fan associations may vary in magnitude and may or may not positively correlate. Additionally, the extent of media consumption can influence the extent to which the perceived performance during events like the World Cup alters preconceived notions. This is due to the fact that the media not only report on mega-events like the FIFA World Cup but also shape and interpret how the tournament and the hosting country are viewed both domestically and globally [10]. The concept of self-categorization has been expanded to incorporate both upward contacts and downward comparisons by Taylor & Lobel, [11] as a means of creating both ingroups and out-groups. This theory holds significant relevance for investigating the underlying issues in the present study.

METHODOLOGY

A. Data Collection

We have collected the data from a previous study [1] that had explored the Fifa 18 world cup. The data collection process for this study involved gathering tweets from users who were fans of the 2018 FIFA World Cup. The tweets were collected in CSV format, with each user's data stored in a separate file. Our dataset is based on twitter. Tweets are based on FIFA 2018 world cup supporter from non-playing countries. There are tweets of 376 users in the dataset. There are 7 teams in the dataset which are Brazil, Argentina, France, Germany, England, Croatia, Portugal. We have worked on the dataset which is collected from twitter. Then used multiclassification process such as Decision tree and random forest. The dataset has tweets of 376 users form non playing countries. we have reduced noisy data. Balanced data using SMOTE. Then we have done word embedding and also tokenization techniques.

B. Preprocessing

We got two kind of data sets: Tweets of 376 supporters in 376 individual files, another file contains user ID, Supporting Team, City where the supporters resides and number of tweets of each supporter. Tweets of 376 supporters has been cleaned by removing stop words, punctuations, and irrelevant words to get a clean and usable dataset. We drop entries with only single characters or stop words. Then all the cleaned tweets of 376 supporters have been saved in a single file.

1. Data Cleaning

The path to the folder containing the tweet data is set. The output path for the cleaned tweets is defined. The cleaning tweet function applies a series of cleaning operations to each tweet. Removes URLs using regular expressions. Removes mentions by matching the '@' symbol followed by alphanumeric characters. Removes hashtags by matching the '#' symbol followed by alphanumeric characters. Removes non-ASCII characters using encoding and decoding techniques. Removes punctuation using the translate method. Removes the ' at the end of the sentence if it exists.

2. Tokenization

Tokenizes the tweet into individual words using the word tokenization function from the NLTK library. Joins the stemmed tokens back into a single string representing the cleaned tweet.

3. Stop words

Removes stop words using a predefined set of English stop words from the NLTK library. Performs stemming using the Porter stemming algorithm from the NLTK library.

4. Data Reading and Concatenation

A list called df_list is created to store the data frames read from each CSV file. For each file in the specified folder:

If the file has a '.csv' extension, the file is read as a data frame with appropriate column names. The 'User Name' column is derived from the file name.

There are some missing values in the dataset which is shown below in figure-1.

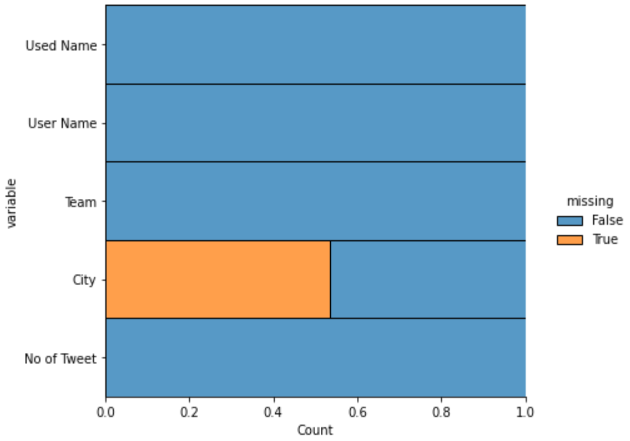

Fig 1: Representation of missing values

1.1 Data Preparation

The Twitter data is loaded from the provided Excel file using pandas library. The column of interest for the classification task, i.e., 'Team' and 'Team_short', are selected. The 'Team' column is mapped to numerical labels representing the supported teams: Brazil (0), France (1), Croatia (2), England (3), Portugal (4), Germany (5), and Argentina (6). The 'Team_short' column is also mapped to numerical labels representing the short codes of the supported teams: BRA (0), FRA (1), CRO (2), ENG (3), POR (4), GER (5), and ARG (6).

1.2 Data Splitting

The prepared data is split into training and test sets using an 80:20 ratio. The features (X) for training data consist of all columns except the last one, which is the target variable.

The target variable (y) for training data contains the numerical labels representing the supported teams.

The features (X) and target variable (y) for the test data are also extracted accordingly.

1.3 Text Vectorization

The text data (tweets) in the training set (X_train) is vectorized using the Count Vectorizer from scikit-learn library.

The vectorizer converts the text data into numerical feature vectors, where each word becomes a feature, and its value represents its frequency in the tweet.

The vectorizer is fitted on the training data and then used to transform both the training and test data into vectorized representations.

1.4 Class Imbalance Handling with SMOTE

The SMOTE algorithm is applied to the training set using the SMOTE class from the imbalanced-learn library.

Among the supporters of all 7 countries, Portugal has only instances of 11 (3%) that makes our dataset imbalanced. The difference of the number of instances with other classes is considerably large. The dataset which has large variance among different class instances is called an imbalanced dataset. Class imbalance problem may lead to poor performance in machine learning prediction. Therefore, we use Synthetic Minority Over-sampling Technique (SMOTE) technique, which is powerful and widely used in imbalanced dataset [3]. We use SMOTE package by using WEKA machine learning toolkit to increase the imbalanced instances. In our study, we apply the SMOTE technique and increase.

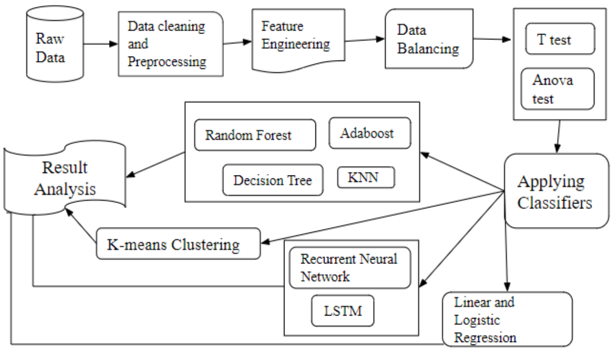

Fig 2: Research Framework of the proposed model

In figure-2, at first, we have collected the raw data of dataset which is collected from twitter at the time of FIFA world cup 2018. Then we have pre-processed both datasets with suitable feature selection methods. We have been used vectorizer and tokenizer techniques. Then we have also used stop word removing techniques. We have done T-test and Anova for to see similarities and differences between the data. Then we have applied our selected deep learning algorithms like LSTM and RNN. Also applied regression models like Logistic regression and Linear regression. Implemented classification model like Random Forest, Decision Tree, KNN, Adaboost. Lastly, we have implemented K-means clustering. When the implementation has been done then we have compared the exiting results with some previous results in this field.

RESULT AND DISCUSSION

We have done our result analysis for the chosen dataset model. A good model's core behavior is that accuracy improves with time and loss reduces with time. Precision, recall, and F1-Score have been calculated as performance evaluation criteria for accuracy. These three criteria are performed to measure the dataset model.

Accuracy: $A=\frac{Tp+Tn}{Tp+Tn+FN+FP}$ Precision: $P = \frac{Tp}{TP+FP} * 100$

Recall: $R = \frac{TP}{TP+FN} * 100$ F1 Score: $F1= \frac{2*P*R}{P+R} * 100$

We Used some popular algorithms (1) Decision Tree, (2) KNN, (3) AdaBoost, (4) Random Forest, (5) LSTM. (6) RNN (7) Logistic and Linear Regression, (8) K-means Clustering.

TABLE I. OVERVIEW OF OUR IMPLEMENTATION BASED ON CLASSIFICATION MODELS.

Model	Accuracy
Decision Tree	0.88
Random Forest	0.32
K nearest neighbour	0.37
Adaboost	0.38

In table-I, we can see that, we have got the highest accuracy from Decision tree among all applied classification models. And lowest from Random Forest.

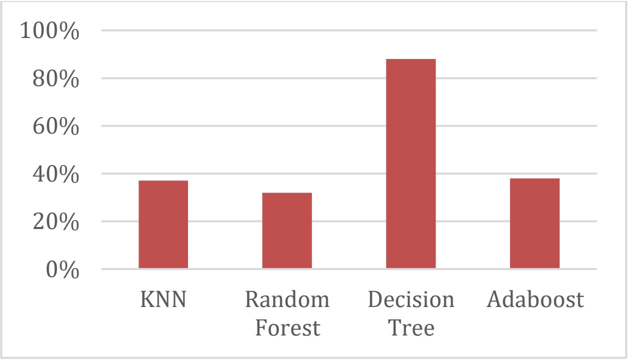

Fig 3: Accuracy of the classification models

In figure-3, we have got the best result from Decision Tree model.

TABLE II. Overview of our implementation based on Deep Learning models.

Model	Accuracy
LSTM	0.31
RNN	0.50

In table-II, we have shown two deep leaning model accuracy result.

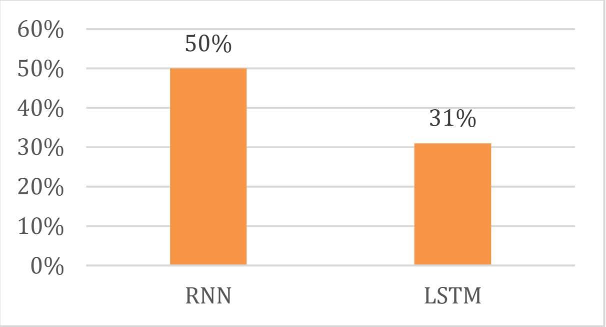

Fig 4: Accuracy of the Deep Learning models

In Figure 4, we can see that, we have got best result from RNN which is 50% and LSTM 31%. Among deep learning models RNN has the best accuracy.

We have done also K-means clustering which result is shown below in the following figures.

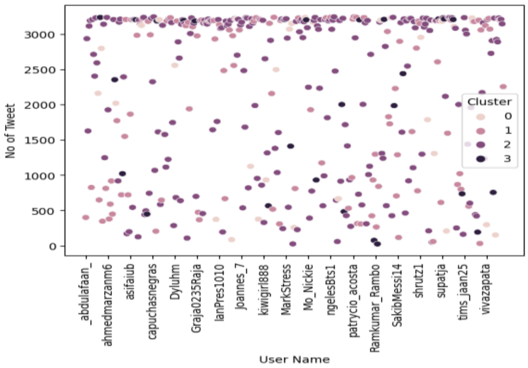

Fig 5: Representation of clusters among some users and tweets.

We can see the clusters of users and their tweets in figure-5.

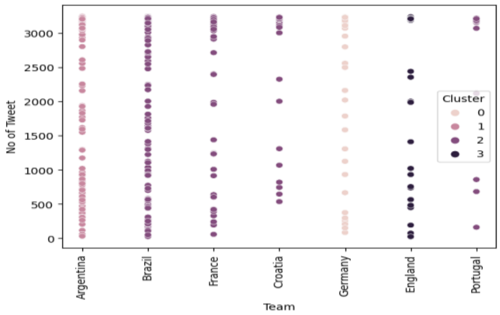

Fig 6: Representation of clusters among Teams and tweets.

We can see the clusters of team and users tweets about supporting teams in figure-6.

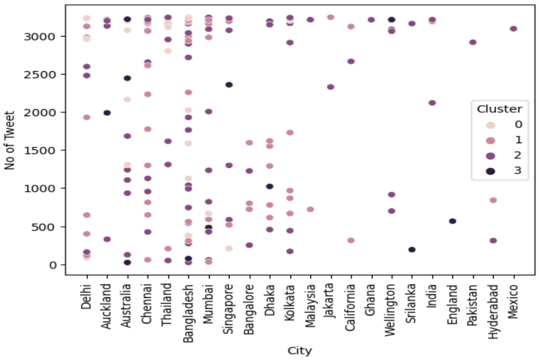

Fig 7: Representation of clusters among City and tweets.

In figure-7, we can see the clusters of city and tweets. We have done Skewness and kurtosis for out data. The skewness of anormal distribution is zero. Negative sign in skewness value indicates that longer tail in the left-hand side. Most of the tweets are between 23 to 2800. The kurtosis of a normal distribution is 3. Negative kurtosis of value less than 3 indicates the tail is short means most of the tweets are concentrated within in a short range.

From the below distribution we can say that it is left skewed and short tailed. The skewness and kurtosis of the distribution are -0.52 and -1.42 respectively.

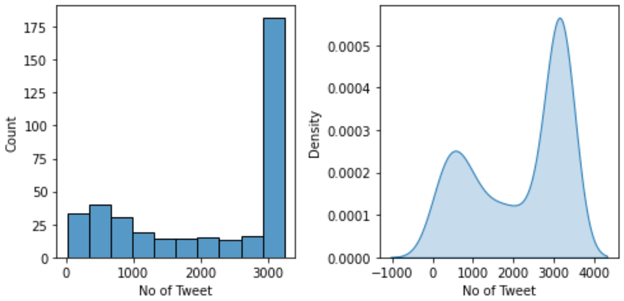

Fig 8: Distribution

Brazil and Argentina are the most popular teams as the supporters from these two countries are 136 and 107 respectively.

Most of the tweets come from Brazil and Argentina supporters which indicates they are the most enthesuiast and crazy fans and love to express their emotion and opinion in the social media.

CONCLUSION AND FUTURE WORK

The prediction of World Cup supporters is influenced by multiple factors, including demographics (age, gender, location), previous attendance, team preferences, and socio-economic indicators. These factors play a significant role in determining an individual's likelihood of becoming a supporter during the tournament. Classification models like Random Forest, Decision Tree, KNN, and AdaBoost have proven to be effective in predicting World Cup supporters. By training these models on historical data that incorporates relevant features, such as demographic information and past behavior, they can accurately classify individuals as potential supporters or non-supporters. In addition to predicting supporter likelihood, regression models can be applied to estimate the number of supporters for different regions or countries during the World Cup. By considering factors like population demographics, historical attendance, and economic indicators, these models provide valuable insights into the expected turnout of supporters. Deep learning models, such as neural networks, have the capability to capture intricate patterns and relationships in the data. By training on large and diverse datasets that include various features related to supporters, these models can provide accurate predictions about the likelihood of individuals becoming World Cup supporters. They can handle both structured and unstructured data, enabling a comprehensive understanding of supporter behavior. Clustering techniques, such as K-means or DBSCAN, can identify distinct segments within the population that are more likely to become World Cup supporters. By clustering individuals based on attributes like demographics, interests, and previous attendance, we can uncover supporter profiles and target specific segments with tailored marketing strategies. Clustering provides valuable insights into the characteristics and preferences of potential supporters.

Overall, the combination of classification models, regression analysis, deep learning models, and clustering techniques allows for a comprehensive prediction of FIFA World Cup 2018 supporters. These approaches consider various factors, capture complex patterns, estimate supporter numbers, and reveal distinct supporter segments, providing valuable insights for decision-making and marketing strategies during the tournament.

Predicting the behavior and preferences of FIFA World Cup supporters can be an intriguing area of research and analysis. While the future of this specific topic is uncertain, here are a few potential directions and considerations for future work in FIFA World Cup supporter prediction:

Future work could focus on developing advanced data analysis techniques to gain deeper insights into supporter behavior. This could involve exploring more comprehensive and diverse datasets, including social media data, ticket purchase patterns, fan forum discussions, and sentiment analysis to capture the sentiment and preferences of supporters. Researchers can explore the use of advanced machine learning algorithms, such as deep learning or reinforcement learning, to improve the accuracy of supporter predictions. These algorithms can analyze large datasets, discover patterns, and make more accurate predictions about fan behavior. Investigating the social network dynamics among supporters can provide valuable insights into the formation of fan communities, fan-driven initiatives, and the spread of information. Future work could focus on analyzing supporter networks, identifying key influencers, and understanding the impact of these networks on supporter behavior. Develop predictive models that can anticipate supporter behavior based on historical data and external factors can be an interesting avenue for future research. These models could consider various factors such as team performance, player statistics, match schedules, cultural and demographic data, and even external events that might influence fan engagement. As new technologies emerge, integrating them into supporter prediction can open up exciting possibilities. For example, incorporating virtual reality (VR) or augmented reality (AR) experiences to simulate fan engagement and measuring the resulting user behavior could provide unique insights into supporter preferences and motivations. With the increasing availability of personal data, it is essential to address privacy concerns and ethical considerations when conducting research on supporter prediction. Future work should aim to develop frameworks and guidelines to ensure the responsible and ethical use of supporter data. These suggestions are speculative and are intended to inspire future research in the field of FIFA World Cup supporter prediction. The actual direction and advancements in this area will depend on technological developments, data availability, and the research community's interests and priorities.

REFERENCES

[1] M. F. Rabbi, M. S. H. Mukta, T. N. Jenia, and A. K. M. N. Islam, *Predicting fans' FIFA world cup team preference from tweets*, vol. 325 LNICST. Springer International Publishing, 2020. doi: 10.1007/978-3-030-52856-0_22.

[2] M. Fan, A. Billings, X. Zhu, and P. Yu, "Twitter-Based BIRGing: Big Data Analysis of English National Team Fans During the 2018 FIFA World Cup," *Commun. Sport*, vol. 8, no. 3, pp. 317–345, 2020, doi: 10.1177/2167479519834348.

[3] M. B. Devlin and A. C. Billings, "Examining the World's Game in the United States: Impact of Nationalized Qualities on Fan Identification and Consumption of the 2014 FIFA World Cup," *J. Broadcast. Electron. Media*, vol. 60, no. 1, pp. 40–60, 2016, doi: 10.1080/08838151.2015.1127243.

[4] D. Miljkovic, L. Gajic, A. Kovacevic, and Z. Konjovic, "The use of data mining for basketball matches outcomes prediction," in *IEEE 8th International Symposium on Intelligent Systems and Informatics*, Sep. 2010, pp. 309–312. doi: 10.1109/SISY.2010.5647440.

[5] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, Feb. 2012, pp. 189–198. doi: 10.1145/2166966.2166999.

[6] C. K. Leung and K. W. Joseph, "Sports data mining: Predicting results for the college football games," *Procedia Comput. Sci.*, vol. 35, no. C, pp. 710–719, 2014, doi: 10.1016/j.procs.2014.08.153.

[7] Y. Bai and X. Zhang, "Prediction Model of Football World Cup Championship Based on Machine Learning and Mobile Algorithm," *Mob. Inf. Syst.*, vol. 2021, 2021, doi: 10.1155/2021/1875060.

[8] H. Fan *et al.*, "Social media toxicity classification using deep learning: Real-world application uk brexit," *Electron.*, vol. 10, no. 11, pp. 1–18, 2021, doi: 10.3390/electronics10111332.

[9] A. S. Gerber and T. Rogers, "Descriptive social norms and motivation to vote: Everybody's voting and so should you," *J. Polit.*, vol. 71, no. 1, pp. 178–191, 2009, doi: 10.1017/S0022381608090117.

[10] J. Maguire, "Invictus or evict-us? Media images of South Africa through the lens of the FIFA World Cup," *Soc. Identities*, vol. 17, no. 5, pp. 681–694, Sep. 2011, doi: 10.1080/13504630.2011.595208.

[11] S. E. Taylor and M. Lobel, "Social Comparison Activity Under Threat: Downward Evaluation and Upward Contacts," *Psychol. Rev.*, vol. 96, no. 4, pp. 569–575, 1989, doi: 10.1037/0033-295X.96.4.569.

GitHub repository link given below:
https://github.com/HasnatAbdullah/FIFA_WC_2018_Supporter_Prediction_Twitter