International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 107-115 www.ijaem.net ISSN: 2395-5252

Social Media Sentiment Analysis and Its Impact on Football Club Performance

Vedran Selak
[1]Student, International Burch University, Sarajevo, Bosnia and Herzegovina

Date of Submission: 01-08-2024 Date of Acceptance: 08-08-2024

ABSTRACT: Social media provides a large pool of user generated content that contains opinions and valuable information on various topics. Sentiment analysis can be used in order to predict the sentiment of user posts and it can be a valuable tool in today's world. This research uses sentiment analysis to analyse the impact of fan sentiment on the performance of the football club. The research performs preprocessing on the tweets directed to the football clubs and merges the tweets with the club performance in the upcoming match. After preprocessing, the data is passed to a Random Forest classifier in order to perform sentiment analysis. The results of the classifier are then analysed and compared to the performance of the football club in their upcoming match. The research showed that positive sentiment before the match increases the chances of success by approximately 1%, leading to the conclusion that the fan sentiment does not impact the performance of the team by a significant margin. The results of this research contribute to the field of natural language processing and provide valuable insights on the topic of Sentiment analysis and fan engagement.

KEYWORDS: Natural Language Processing, Sentiment Analysis, Social Media, Text Classification, Preprocessing, Engagement, Sport, Football.

I. INTRODUCTION

With the rise of social media the user generated data over the internet has risen exponentially. Social media platforms enable users to share their knowledge, opinions, and experiences to the public on various topics. Analysing this vast amount of data manually poses a difficult task and automated techniques are required to make the process viable. Techniques such as Sentiment analysis help in analysis of such data [1]. Sentiment analysis is used to extract sentiment from textual data and is a part of a broader field of study called Natural Language Processing (NLP) [2]. Sentiment analysis requires detection, extraction and classification of sentiment from different sources and structures, such as blogs, news, social networks, forums, etc. [3] The research focuses on performing sentiment analysis on fan tweets, mentioning the Football clubs from the English Premier League, analysing its impact on the performance of the club in the upcoming match. The aim of the research is to understand, does fan sentiment leading to the match influence the performance of the club in that match. By analysing the emotional engagement and sentiment expressed by fans leading up to the match, the research aims to discover whether the collective mood of the fan base can influence players' morale, team dynamics, and finally the outcome of the match. The following is the research question that authors are aiming to answer within this research: Does the sentiment of the fans influence the performance of the football club? The above mentioned research question, once answered, will provide a meaningful contribution to the field. The main focus of the question is to see if fans can improve the performance of their club by being more positive and supportive towards the players and the club. The rest of the paper is organised as follows. The next section describes the related work that has been done. The methods used in the research are described in the third section. The fourth section describes the data and the retrieval process. The findings will be described in the fifth section and the conclusion section brings this research to a close.

II. LITERATURE REVIEW

Schumaker et al., [4] conducted research focused on predicting the outcome of football matches by using the sentiment contained in the tweets. The Authors collected tweets for the twenty clubs in the English Premier League, analysed the crowdsourced sentiment and applied the predictions to a wagering decision system. The research compared the results of 8 sentiment based models with the 1 odds-favourites based model. The research showed that the odds-favourite model outperformed the 8 sentiment based models in accuracy, however 5 of the 8 sentiment based models overall earned more money from their bets. A similar study was performed by Godin et al. [5]. The authors focused on predicting the winner of football games using methods that use Twitter volume, the sentiment and the score predictions made by twitter users, in order to beat the bookmakers. For the paper the authors used 200 games from the English Premier League and were able to beat the bookmaker by combining multiple methods together. The sentiment model's overall accuracy was 52%, while combining multiple methods by using early fusion obtained the accuracy of 68%. Furthermore, Souza et al. [6] conducted research in order to develop a football prediction system, while considering diverse parameters in order to obtain accurate results. They also decided to perform sentiment analysis on tweets regarding the English Premier League matches and football clubs. The authors used Logistic regression, SVM and Random forest to predict the matches, the authors also performed analysis on the football player facial expressions and the fan sentiment. The sentiment analysis was used to understand the main emotions experienced by fans through the match, enabling stakeholder and players to better understand and connect with their fans. The paper conducted by S. Aloufi and A. E. Saddik [7], focused more on the process of sentiment analysis and provided a detailed analysis of the performance of different algorithms and feature selection processes. Authors used Twitter (now X, during the paper it will be referred to as Twitter) posts during the Champions League and the FIFA 2014 World Cup in order to generate a dataset. The paper compares three learning algorithms (Support Vector Machine, Multinomial Naïve Bayes and Random Forest), as well as three models for creating features (Bag-of-Words, Part-Of-Speech and Existing Sentiment Lexicons). The research showed that the SVM algorithm combined with the BOW model showed robust and consistent performance in comparison with MNB and RF. Similarly, Aloufi et al. [8] conducted a sentiment analysis on tweets related to the UEFA Champions League 2016/2017 season. The authors generated a football specific sentiment lexicon that in combination with other lexions boosted performance by approximately 6%. The authors used the SVM algorithm and compared the performance of different combinations of lexions as well as Uni-gram and Bi-gram BOW feature models. The research showed that Uni-gram models perform better than Bigram models and the performance is better if the required word frequency is lower (minimum 5 times). Another study conducted by Singh et al. [9] presents a comparison of different feature evaluation and classification techniques for sentiment analysis on social media data. The compared feature extraction techniques were BOW, POW and feature hashing (HT) and the classifiers were Naïve Bayes, SVM, RF and Linear Regression. The authors used the Twitter sentiment corpus and the Stanford dataset for obtaining the tweets. The research showed that the SVM algorithm combined with POS feature extraction gave the best results. Wunderlich F. and Memmert D. conducted a study that focuses on in-play forecasting, by extracting information from two million tweets by applying sentiment analysis [10]. The authors performed preprocessing by removing the URLs, mentions, punctuation, hashtags, emoticons, characters and digits. The cleaned tweets were analysed by three different sentiment analysis methods and the average score of positivity and negativity was assigned to each tweet. The results showed that in-play forecasting is very challenging and does not improve forecasting accuracy when compared with pregame information. The analysis also showed that the fans' sentiment towards the match gets more negative as the time goes by, meaning that the sentiment is decreasing over the course of the match. Table 1 shows the summary of the works listed above. It focuses on the results obtained by the different studies and the methods used in order to obtain those results. Extracting the best results is difficult since many studies focus on different aspects within the same domain of research. Each study provides insightful information regarding sentiment analysis and match prediction in football matches.

STUDY	METHOD	RESULTS
Schumaker et al. [4]	Predicting matches using sentiment analysis performed by OpinionFinder.	The authors build 8 sentiment models. The best model focused on subjective negative tweets, obtaining 50.49% accuracy
Godin et al. [5]	Predicting matches usingsentiment	The model had 52% overallaccuracy and

	analysis performed with SVM classifier and vectorizes with TF-IDF.	it peaked with 56% accuracy with one data batch.
Souza et al. [6]	Used sentiment analysis to assign emotion to matches and used the data to predict the outcome of the matches with SVM and RF.	The authors mention that the best algorithm was the SVM algorithm, however they didn't mention the exact metric numbers.
S. Aloufi and A. E. Saddik [7]	Authors compared compares three algorithms: SVM, MNB and RF, as well as three models for creating features: BOW, POS and Existing Sentiment Lexicons.	The authors created a large number of models combining machine learning algorithms with one or multiple feature extraction methods. Overall, the SVM algorithm had the best accuracy score and peaked at 59% accuracy (SVM with BOW+POS+General Lexicon).
Aloufi et al. [8]	The authors performed sentiment analysis using SVM and experimenting with different feature extractions and lexicons.	The authors obtained 85% accuracy using SVM with BOW+Lexicons+linguistics (number of uppercase words, exclamation and question marks).
Singh et al. [9]	Authors performed sentiment analysis while comparing SVM, NB, Linear Regression and RF algorithms in combination with POS, BOW and HASS tagging feature extraction techniques	The best combination for sentiment analysis was the SVM classifier and the POS feature extraction technique, obtaining 83.27% accuracy.
Wunderlich F. and Memmert D. [10]	The authors focused on ingame forecasting using Logistic Regression, RF and other statistical methods (UNI, FRQ).	The research showed that RF and Logistic Regression were unable to outperform pregame betting odds.

TABLE 1. LITERATURE REVIEW SUMMAY

III. METHODOLIGY

Football clubs all around the world use social media to influence and communicate with their fans. This generates a lot of useful data that can be analysed to potentially have an edge over the competition. This research attempts to understand the sentiment of the fans and the influence of the sentiment on the club's performance. It is achieved with Sentiment Analysis of the tweets that are mentioning the football clubs, which are afterwards compared with the club performance in their upcoming match. Before building a model for the sentiment analysis, the tweets need to be preprocessedin order to enable the model to be effectively trained. The preprocessing consists of multiple steps and the first step is to remove the mentions, hashtags, links, numbers, punctuation marks and non-alphanumeric characters, as well as convert each tweet to lowercase. With this step we are able to remove unnecessary content that has a low chance of providing value to the sentiment analysis model. In order to demonstrate the process the following tweet "James Justin deserves a mention here also. Put a great shift in for Leicester on Sunday and should get a run of games to display his ability, with Pereira out injured. #LCFC" is preprocessed as an example, step by step. Table 2 shows the tweet's content before and after applying the above described methods.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 107-115 www.ijaem.net ISSN: 2395-5252

TIME	TWEET
Before	James Justin deserves a mention here also. Put a great shift in for Leicester on Sunday and should get a run of games to display his ability, with Pereira out injured. #LCFC
After	jamesjustin deserves a mention here also put a great shift in for leicester on sunday and should get a run of games to display his ability with pereira out injured.

TABLE 2. BEFORE AND AFTER CLEANING.

Another important step of preprocessing is removing stopwords and non english words. Stopwords often appear in natural language text, however they provide very little information about the part of the text they belong to. By removing stop words we are able to decrease the noise in the data and improve the performance of the sentiment analysis [11]. The stopwords for the research were obtained via NLTK's stopwords toolkit. The NLTK toolkit provides a vast amount of corpuses, lexical resources and NLP libraries that is used throughout the whole research implementation. Table 3 demonstrates the changes applied to the tweet.

TIME	TWEET
Before	jamesjustin deserves a mention here also put a great shift in for leicester on sunday and should get a run of games to display his ability with pereira out injured
After	mention also put great shift get run display ability pereira injured

TABLE 3. BEFORE AND AFTER REMOVING STOPWORDS AND NON-ENGLISH WORDS.

Furthermore, in order to reduce different variants of a specific term to a single one, lemmatization is performed on the cleaned tweets. This step makes it easier to perform vectorization and transform the fan interactions in data that can be fed to machine learning classification algorithms. Lemmatization is used in order to reduce and simplify the vocabulary size and the variety of the text, enabling better analysis of the sentiment. Lemmatization is performed with the NLTK's WordNetLemmatizer. Table 4 shows the tweet's contents before and after applying Lemmatization.

TIME	TWEET
Before	mention also put great shift get run display ability pereira injured
After	mention also put great shift get run display ability pereira injure

TABLE 4. BEFORE AND AFTER APPLYING LOEMMATIZATION.

With lemmatization the data cleanup process is finalised, in summary it consists of three steps: cleaning the tweets of unwanted characters, removing the stopwords and non english words, and performing lemmatization on each word. The Figure 1 shows the code used for each step, where each step is extracted into a function.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 107-115 www.ijaem.net ISSN: 2395-5252

```python
# cleaning tweet text
def clean_tweet(tweet):
    # remove mentions, links, hashes
    tweet = re.sub(r"([@|#][A-Za-z0-9_]+)|#|http\S+|www\S+", "", tweet)
    # remove numbers
    tweet = re.sub("[0-9]", '', tweet)
    # remove punctuation
    tweet = re.sub('[()!?\"":]', ' ', tweet)
    # remove newline
    tweet = re.sub('\n', ' ', tweet)
    # remove non-alphanumeric characters
    tweet = re.sub('[^\w]', ' ', tweet)
    # lowercase
    tweet = tweet.lower()
    return tweet

# eliminating stopwords and unkown words
words = set(nltk.corpus.words.words())
def extract_words(tweet):
    extracted = [word for word in tweet.split() if word.lower() not in stopwords.words('english') and word.lower() in words]
    return extracted

# extracting word lemmas
def lemmatize(tweet_list):
        lem = WordNetLemmatizer()
        lemmatized_tweet = []
        for word in tweet_list:
            lemmatized_text = lem.lemmatize(word,'v')
            lemmatized_tweet.append(lemmatized_text)
        return lemmatized_tweet
```

FIGURE 1. PYTHON CODE FOR EACH PREPROCESSING STEP

Before we start building the model the cleaned data is converted into a data form that is meaningful to the machine, this is achieved with vectorization. For executing vectorization we are using the CountVectorizer from the Scikit learn library. With CountVectorizer we have transformed our data into a set of features, creating a Bag Of Words model. In this research the algorithm used for sentiment prediction is the Random Forest (RF) [12] algorithm. RF is a popular machine learning algorithm that is used to create prediction models. The RF algorithm consists of multiple decision trees that form a forest. The decision trees are easy to use and are an intuitive method for predicting results. However, the decision tree method often performs poorly when it's working with complex data. In RF many decision trees are created using randomly selected data and random set of features for creating predictions. The results of each decision tree are aggregated in order to give a final prediction [13]. This makes the RF better at creating predictions on datasets that have a high number of features, this certainly applies to this research since sentiment analysis with the Bag Of Words model is very feature heavy. However, before using the data to train the RF classifier it needs to be split up into two subsets. One subset is used for training the classifier and it consists of 80% of the actual data, while the other subset is used for testing of the trained model and it contains the remaining 20% of the original dataset size. For training the model the RF classifier contains 100 decision trees, meaning that results from 100 decision trees are aggregated in order to provide a final answer. After the classifier has been trained, it is then used to predict the sentiment of the test subset and the prediction of the sentiment is then mapped to the original tweet that contains the results of the upcoming match. With this new dataset we are able to perform analysis on the amount of positive and negative sentiment leading to certain results in the upcoming matches. The whole process of data processing and performing sentiment prediction on the data is shown via a flowchart illustrated in Figure 2.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 107-115 www.ijaem.net ISSN: 2395-5252

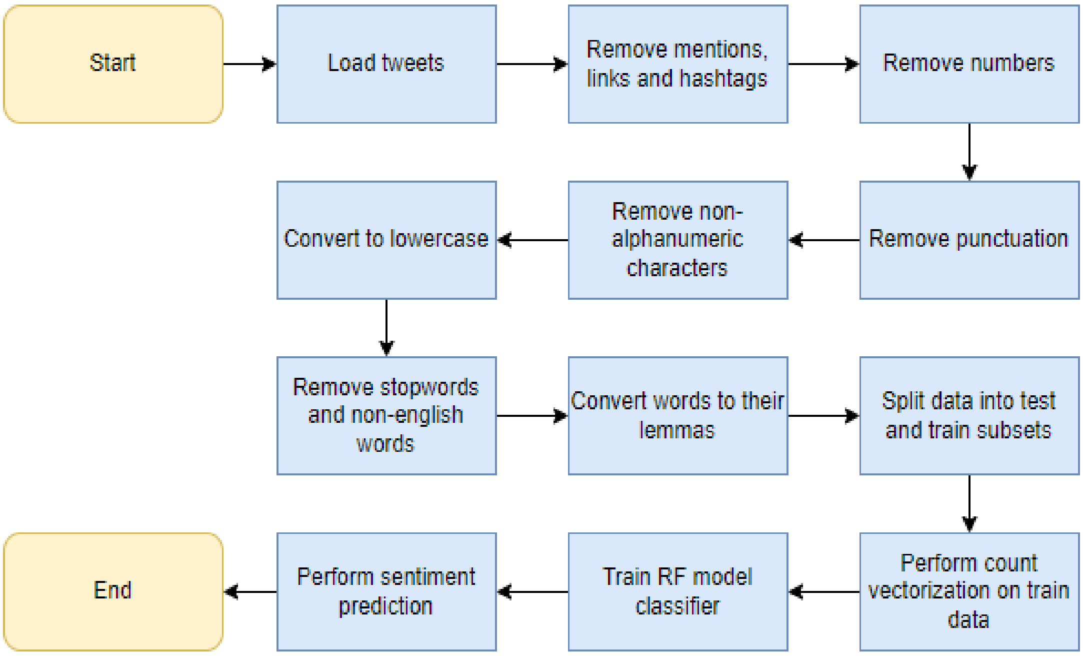

FIGURE 2. PREPROCESSING AND SENTIMENT PREDICTION FLOWCHART.

IV. DATA

The data for this research is obtained from kaggle and the research consists of two datasets. One dataset contains the tweets mentioning the football clubs while the other dataset contains the result of matches of those football clubs. Both datasets focus solely on the football clubs and the matches that occur within the English Premier League. The two datasets were used in conjunction in order to generate a single dataset which could be used for the final goal of this paper. Below we explain the contents of both datasets, represent those values in two tables and finally explain the method used for combining the two datasets in order to analyse the impact of the fan sentiment on the club performance. The dataset containing the tweets [14] collected the data from the Twitter API and it consists of approximately 1 million tweets from the 20th of July till the 19th of September 2020. The dataset contains 15 columns, however for the purpose of this research only three columns were used: created_at, text, partition_1 (will refer to it as "club") and polarity. The club column contains the football club names at which the tweet is directed to. This column is required in order to map the tweet to its corresponding club. The polarity column is the sentiment computed from the tweet and it is used in order to train the RF classifier using supervised learning. The text column contains the actual tweet made by the user, this column is the main focus of the research paper. The final column from this dataset is created_at, this column enables us to find the nearest upcoming match of the club the tweet is directed at. The dataset with example data is shown in Table 5. The dataset containing the match results [15] covers the match history of all English Premier League matches from the start of 1992 till the final week of the 2021- 2022 season. This is approximately 12 thousand matches each containing 8 columns of related value. For the purpose of this research we used the Home, Away, Date and FTR columns. The Home and Away columns hold the names of the football clubs participating in the match. The FTR column contains the match results, it consists of three values H (Home team won), A (Away team won), D (a draw, even score). The Date column is used in order to map the match to the tweet. The dataset with example data is shown in Table 6.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 08 Aug. 2024, pp: 107-115 www.ijaem.net ISSN: 2395-5252

NAME	TYPE	EXAMPLE
created_at	string	2020-09-23 18:20:43
text	string	HE WORKS WHEN HE WANTS, HE WORKS WHEN HE WAAAANTS, @HLTCO HE WORKS WHEN HE WANTS #CPFC
club	string	CrystalPalace
polarity	number	1

TABLE 5. TWEETS DATASET EXAMPLE ROW.

NAME	TYPE	EXAMPLE
Home	string	Leeds United
Away	string	Liverpool
Date	string	2021-09-12
FTR	string	A

TABLE 6. MATCHES DATASET ROW.

The sentiment analysis was performed on the cleaned text column of the tweets dataset, the results were then mapped back to the original row that contains the tweet and the other columns. The final step before analysis is to join the two datasets. Since this research focuses on analysing the impact of the sentiment on future matches, each tweet was mapped to the nearest upcoming match of that club. In order to do that we added a new column to the tweets dataset (upcoming_match) by observing the dates of the matches and finding the first match of the appropriate club that has a date larger than the tweet's date. Once the match is found the FTR is observed and based on the FTR and the location of the club (Home or Away), the results are documented as 2 (for win), 1 (for draw) and 0 (for lose). Unfortunately, more than 2 thirds of the tweets were made during the time between two football seasons. These tweets couldn't be assigned to any match, since the time between the tweet creation and the match is too large. Therefore those tweets were removed from the research data, leaving us with 297 988 tweets to conduct sentiment analysis and analyse the results. With this being done, the final dataset contains all the required data that is needed to perform the analysis.

V. RESULTS

The analysis was conducted by observing the percentages of fan sentiment when the upcoming match is a win and the fan sentiment percentages when the upcoming match is a loss. The data showed that before the upcoming win the fan sentiment was 72.1% positive with the remaining 27.9% being negative. When the upcoming match is a loss the data shows a very slight dip in positive sentiment, 70.8% positive sentiment and 29.2% negative. Both charts are shown in Figure 3 and 4.

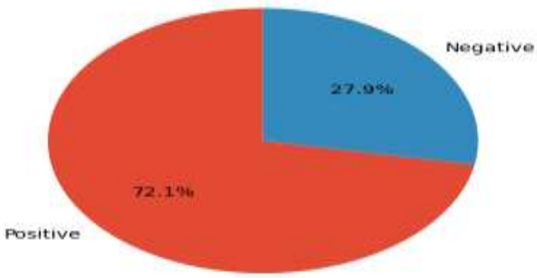

FIGURE 3. UPCOMING WIN SENTIMENT.

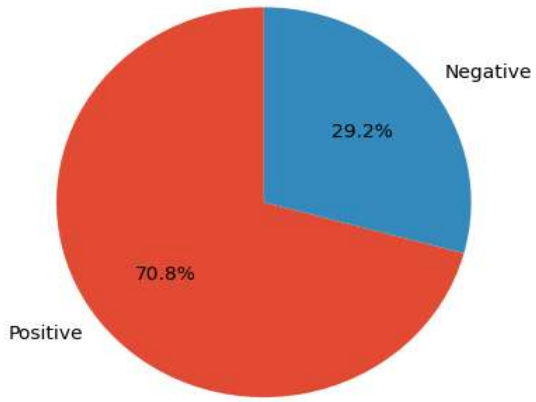

FIGURE 4. UPCOMING LOSS SENTIMENT

The research shows that the fan sentiment has a very insignificant impact on the club results in their upcoming match. The 1.3% difference in positive sentiment is not enough in order to claim that by being more positive towards the team and the club that the club will perform better in the future. Furthermore, we analysed the win percentages per each club in the dataset. In the dataset there are 9 football clubs that during the timeframe of the available tweets have both won and lost a game. This allows us to compare the percentages of the fan sentiment leading to a win and a loss. The data shows that 3 clubs had a larger percentage of positive sentiment leading up to a win, compared to the positive sentiment percentage leading up to a loss. The opposite was true for 4 clubs, while 2 clubs didn't show any variation in percentages. The RF model, used for sentiment prediction, performed with 83% accuracy. When observing the precision and the recall of the model, it shows that the model is much better at predicting the positive sentiment. The precision for tweets with negative sentiment was 0.56, while the precision for positive sentiment was 0.94. The recall shows less variation, with 0.79 on negative sentiment and 0.84 on positive sentiment. The research leads to the conclusion that the fan sentiment does not impact the performance of the club. Meaning that football fans can't improve the performance of their favorite clubs by deliberately interacting with them in a positive manner.

VI. CONCLUSION

This research focused on performing sentiment analysis and analysing the influence of fan sentiment on the performance of the football club. It describes the techniques used for obtaining data and cleaning the tweets in order to optimise the performance of the classifier. As a result, it is shown that a fan sentiment does not influence the performance of the club in the future matches. The sentiment ratio is consistent whether the upcoming match results in a win or a loss. The findings of this research contribute to the field of natural language processing and contribute to understanding the influence of fan engagement on performance of the football clubs. Overall, Sentiment analysis of social media posts can improve understanding of the public opinion on different topics and create valuable insight from the vast amount of user generated data. These insights can be used by various businesses, organisations and individuals in order to improve marketing strategies, customer satisfaction and customer engagement. The limitations of this research are on the amounts of data used for analysis and the time frame of the tweets. For future work authors would recommend to observe the performance of the clubs throughout the whole season and the sentiment of the fans changing depending on the club performance, and vice versa. Additionally, since football tweets are very context heavy it would be beneficial to build a football specific vocabulary and perform the sentiment analysis using it in order to increase classifier's accuracy.

REFERENCES

[1]. Singh, N. K., Tomar, D. S., &Sangaiah, A. K.: Sentiment analysis: a review and comparative analysis over social media. In Journal of Ambient Intelligence and Humanized Computting. vol. 11, issue 1, pp. 97–117. Springer Science and Business Media LLC (2018). doi: 10.1007/s12652-018-0862-8

[2]. Chauhan, P., Sharma, N., & Sikka, G.: The emergence of social media data and sentiment analysis in election prediction. In

Journal of Ambient Intelligence and Humanized Computting. vol. 12, issue 2, pp. 2601–2627. Springer Science and Business Media LLC (2020). doi: 10.1007/s12652-020-02423-y

[3]. Ramírez-Tinoco, F. J., Alor-Hernández, G., Sánchez-Cervantes, J. L., Olivares-Zepahua, B. A., & Rodríguez-Mazahua, L.: A Brief Review on the Use of Sentiment Analysis Approaches in Social Networks. In Advances in Intelligent Systems and Computing. pp. 263– 273. Springer International Publishing (2017). doi: 10.1007/978-3-319-69341-5_24

[4]. Schumaker, R. P., Jarmoszko, A. T., & Labedz, C. S., Jr.: Predicting wins and spread in the Premier League using a sentiment analysis of twitter. In Decision Support Systems. vol. 88, pp. 76–84. Elsevier BV (2016). doi: 10.1016/j.dss.2016.05.010

[5]. Fréderic Godin, Zuallaert, J., Vandersmissen, B., Neve, W. D., & Walle, R. V. D.: Beating the Bookmakers: Leveraging Statistics and Twitter Microposts for Predicting Soccer Results. KDD Workshop on Large-Scale Sports Analytics. Sydney, Australia (2014). doi: 10.13140/2.1.2168.0000

[6]. Souza, N.J.D., Samrudh, H.N., Gautham, S., Shaman Bhat, B.U., Nagarathna, N.: Football Game Analysis and Prediction. In: Reddy, V.S., Prasad, V.K., Wang, J., Reddy, K.T.V. (eds) Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing. vol 1325. Springer, Singapore (2021). doi: 10.1007/978-981-33-6912-2_16

[7]. Aloufi, S., & El Saddik, A.: Sentiment Identification in Football-Specific Tweets. In IEEE Access. vol. 6, pp. 78609–78621. Institute of Electrical and Electronics Engineers (IEEE) (2018). doi: 10.1109/access.2018.2885117

[8]. Aloufi, S., Alzamzami, F., Hoda, M., & El Saddik, A.: Soccer Fans Sentiment through the Eye of Big Data: The UEFA Champions League as a Case Study. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 244-250. IEEE, Miami (2018). doi: 10.1109/mipr.2018.00058

[9]. Singh, N. K., Tomar, D. S., &Sangaiah, A. K.: Sentiment analysis: a review and comparative analysis over social media. In Journal of Ambient Intelligence and Humanized Computing. vol. 11, issue 1, pp. 97–117. Springer Science and Business Media LLC (2018). doi: 10.1007/s12652-018-0862-8

[10]. Wunderlich, F., &Memmert, D.: A big data analysis of Twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football? In Social Network Analysis and Mining. vol. 12, issue 1. Springer Science and Business Media LLC (2021). doi: 10.1007/s13278-021-00842-z

[11]. Sarica, S., & Luo, J.: Stopwords in technical language processing. In D. R. Amancio (Ed.), PLOS ONE . vol. 16, issue 8, p. e0254937. Public Library of Science (PLoS) (2021). doi: 10.1371/journal.pone.0254937

[12]. Breiman, L.: Random Forests. In Machine Learning. vol. 45, issue 1, pp. 5–32. Springer Science and Business Media LLC (2001). doi: 10.1023/a:1010933404324

[13]. Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E.: A comparison of random forest variable selection methods for classification prediction modeling. In Expert Systems with Applications. vol. 134, pp. 93–101. Elsevier BV. (2019). doi: 10.1016/j.eswa.2019.05.028

[14]. EPL Teams - Twitter Sentiment Dataset. Kaggle. https://www.kaggle.com/datasets/wjia26/epl-teams-twitter-sentiment-dataset, last accessed 2024/02/22

[15]. Premier League matches 1993-2023. Kaggle. https://www.kaggle.com/datasets/evangower/premier-league-matches-19922022, last accessed 2024/02/22