# Literature Review: Machine Learning techniques for predicting football/soccer match outcomes using social media indicators.

## 1. Introduction

Several machine learning models have been used in the past to predict the outcome of matches (i.e., wins, losses, and draws). These predictions have been useful for bookmakers, sports betting platforms, and gamblers who bet on match outcomes. Sports fans rely on expert insights for match predictions for upcoming games. Players and team management use prediction factors to plan their team line-up and improve tactics and strategies ahead of the games.

There has been a growing body of research on machine learning techniques being applied in predicting football matches and their importance in team management and auxiliary services dependent on sports outcomes. This study builds on the study and research of Reed and Donoghue (2005), Buursma (2010), and Huang and Chang (2010), who used various machine learning models to predict match outcomes using sports features, including but not limited (match venue, rest, team's position in the league, shots on goal, corner kicks, and possession). However, the authors concluded that other variables, such as motivation and injury, were not included, and recommended that complex recognition technology be used in future studies. The literature review will examine other variables beyond sports features that can be used to predict match outcomes, particularly social media content.

## 2. Theoretical Background

### <u>Machine learning in sports prediction</u>

Machine learning models have been used in sports prediction throughout the years, providing useful insights into sports performance and strategy. The English Premier League is the most-watched sports league globally, with an estimated audience of 4.7 billion. (https://en.wikipedia.org/ wiki/Premier League). Having a predictive system has significant economic value to teams and fans. The body of research on predictive systems for soccer matches has used feature engineering to extract meaningful features and apply machine learning techniques to achieve results. Football prediction is a multi-class classification problem with the results falling in one of three classes (Home Win, Away Win, Draw), Baboota and Kaur (2019). Previous studies have explored goal-based predictions, with the aim of predicting goals scored by each team, while other studies looked at results-based approaches, predicting wins, losses and draws.

The growing accessibility of sport-related data online, coupled with the rise of online sports betting, has increased engagement from sports professionals and former athletes who share analyses and predictions on social media platforms. Bettors frequently utilise these insights to inform their decisions, while fans reference them when considering whether to view a particular game. Joseph, Fenton & Neil (2006) found that incorporating expert knowledge into a Bayesian Network model has a stronger performance in predicting accurate results. The authors used the Bayesian Network to predict the results of Tottenham Hotspur for the period 1995-1997. The authors focused on a results-based approach and demonstrated that the Bayesian Network machine learning technique outperformed techniques such as k-nearest neighbour. The limitations of their study were that it focused on one team for a particular period and only focused on features related to in-game play.

Owramipur, Eskandarian, and Mozneb(2013) explored other features, including weather conditions, player motivation or psychological state, and injuries of main players. Using the Bayesian Network, the authors reported a high accuracy of match prediction of 92%. Again, the system was modelled for one team over 20 matches.

## Use of social media as a predictive tool

Godin, Zualaert, Vandersmissen, De Neve and Van de Walle (2014) took a different approach. The authors made use of the collective knowledge from Twitter. Sentiments from tweets were used in predicting match results and decisions on wagering for bettors. Compared to Joseph et al (2006), who used domain experts, Godin et al. Al (2014) proved that using crowd-sourced experts yielded better results.

Kampakis & Adamides (2014) compared predictive models using data mined from Twitter versus predictive models that used historical data and simple football statistics versus predictive models that combined both Twitter and historical data. Their study focused on the prediction of the outcome of a game in the English Premier League as being a win for the home team, a win for the away team or a draw. The features in the dataset included home team features, away team features, and response variables. The features used had three variations: bag-of-words (Twitter dataset), historical features, and combined (bag-of-words plus historical features). Naïve Bayes, Random forests, Logistic regression and SVM were models used to test out the three models.

The results showed that Naïve Bayes was the best classifier for the historical data model, while random forest provided more accurate results for the Twitter model and for the combined model, which used both Twitter datasets and historical data, Random Forest was considered the best.

Rathan et al. (2018) used Twitter APIs for developers to extract tweets from the Twitter platform to be used to predict football match outcomes. The SVM algorithm was used to perform the sentiment analysis on the tweets. In addition, the authors considered the odds-on favourite, players' and teams' current form to improve the outcome of the prediction.

Kabakuş et al. (2018) collected over thirty-eight million tweets from seven million unique Twitter users during the 2018 FIFA World Cup. The study aimed to develop a prediction system that evaluated teams that qualified for the FIFA World Cup 2018, through their squad and performances in the early stages of the tournament, with an aim to predict the match results further on in the tournament. The machine learning model used novel features based on social media analysis alongside statistical features. The novel features used were (1) the number of Ballon d'Or candidates the national team has, and (2) the ratio of the number of tweets posted for the team to the population of the country which the team represents. Nine machine learning algorithms were used: BayesNet, Naïve Bayes, Naïve Bayes Multinomial, SVM (Support Vector Machines), kNN (k-Nearest Neighbour), Random Forest, Random Tree, Multilayer Perceptron, and Logistic Regression. The Multilayer Perceptron predicted with 8% accuracy the results of the elimination stage of the World Cup. In the round of 16, the SVM algorithm predicted with 87.5% accuracy the match outcomes. The authors stated that, to their knowledge, no recorded work provides better accuracy in terms of predicting match results. The authors proposed future work in applying the proposed system of matches to the UEFA European Football Championship 2020 and the FIFA World Cup 2022 to prove the accuracy of predicting match results for other competitions.

Miranda-Peña et al. (2021) gathered tweets from 54 matches played during the 2019/2020 English Premier League soccer season, starting from week 38 of 2019 through to week 8 of 2020. The authors opted to use a unique methodology for mining sports sentiment in social networks by engineering centrality measures. These were used as candidate features to train the Machine Learning Model. To classify a match as a win, draw or loss at home, Support Vector Machines were used. The authors concluded that the study gave satisfactory results, given that it ignored statistics but instead used the knowledge of fans' comments collated into a historical database to score a team's polarity and generate predictions before the game began.

Le, Ferrara, and Flammini (2015) investigated the design, implementation and validation of a machine learning framework to predict rare outcomes in soccer matches by analysing the sentiment of Twitter conversations related to these games. The choice of rare outcome matches was based on the profitability of the bets placed upon them, as one of the results had very high odds. Tweets were extracted from supporters of teams from six different competitions, including the 2014 FIFA World Cup. Tweets were extracted six hours before the commencement of the match on the average mood of supporters and used the discrete representation to train the machine learning classifier. The authors managed to prove that the system could achieve an accuracy of around 80% in predicting rare outcomes.

## 3. Future Direction

Football is by far the most watched and popular sport in the world, with more than 3 billion fans worldwide (*Le et al.,2015).* The hashtag #WC2022 generated 147 billion impressions globally, proving that conversations surrounding football are popular and engaging. Social media has become a useful and effective source of information that, when paired with football and in-game statistics, powerful prediction models can be built. The various authors have proved that social media, especially Twitter, can be used to predict match outcomes. It is important to note that the use of hashtags on Twitter makes it a lot easier for conversations on a specific topic to be tracked. For future studies, it would be beneficial to look at other social media platforms that use hashtags to tag similar content or discussions. With the advent of TikTok, which has become a popular social media platform boasting an estimated 1.12 billion active users (Datareportal, n.d), this could be a potential area of sentiment analysis studies on match predictions.

It is worth noting that the studies focused predominantly on English language tweets, limiting the variety of tweets that could be used in the sentiment analysis. Given that football is a global sport, multilingual and cross-cultural analysis should be considered. Alhadlaq & Alnuaim (2023) explored the emotional and sentimental cultural traits of Arabic and Hispanic viewers of the 2022 FIFA World Cup Match between Argentina and Saudi Arabia through tweets. The aim was to determine whether cultural diversity is prevalent in online interactions.

## 4. Challenges

There are few studies on other forms of social media besides Twitter for football match predictions, making it difficult to evaluate the effectiveness of the studies across multiple platforms. Twitter discussions can often become nosey with bots that can cause a class in balance between positive and negative sentiments. This could result in the prediction models being biased due to the proliferation of bots tweeting rapidly, viewpoints that may not necessarily reflect crowd-sourced knowledge.

Sentiment analysis alone cannot be used in predicting match outcomes; in-game statistics, player motivations, injuries, player line-ups and team strategy all have a bearing on the ultimate performance of the team. Kabakuş et al. (2018) examined this, in part taking both Twitter sentiments and player statistics.

## 5. Conclusion

The findings presented highlight that there is significant potential in leveraging social media platforms such as Twitter for predicting match outcomes through sentiment analysis. The popularity of football worldwide draws massive engagement on platforms such as Twitter and offers invaluable data through crowd-sourced knowledge and insights. When combined with player statistics and historical data, the machine learning models can achieve a notable success rate.

Sentiment analysis cannot solely be the predictor of match outcomes. Other critical factors need to be considered, such as player motivation, injuries, team strategies, and in-game statistics. Furthermore, the growing concern about the noise generated on social media through bots could introduce biases, affecting the overall reliability of the model.

Current studies are limited to Twitter on match outcome predictions; further studies could explore new social media platforms such as TikTok to diversify the data sources. Given that football viewership is cross-cultural, conducting multilingual and cross-cultural analyses will capture the global nature of football fans and deepen the insights generated through crowdsourced data. By integrating these advancements with robust in-game statistics, researchers can build more comprehensive prediction models, potentially transforming the forecasting of football results into a more accurate and versatile tool.

# References

1) Reed, D. and O'Donoghue, P., 2005. Development and application of computer-based prediction methods. International Journal of Performance Analysis in Sport, 5(3), pp.12-28.
2) Buursma, D., 2010. Predicting sports events from past results. In 14th Twente Student Conference on IT (Vol. 21).
3) Huang, K.Y. and Chang, W.L., 2010, July. A neural network method for the prediction of the 2006 World Cup football game. In The 2010 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
4) Wikipedia contributors. (2024). Premier League. Wikipedia. Available at: https://en.wikipedia.org/wiki/Premier_League [Accessed 27 Jul. 2025].
5) Bunker, R., Yeung, C. and Fujii, K., 2024. Machine learning for soccer match result prediction. In Artificial Intelligence, Optimisation, and Data Sciences in Sports (pp. 7-49). Cham: Springer Nature Switzerland.
6) Baboota, R. and Kaur, H., 2019. Predictive analysis and modelling football results using a machine learning approach for the English Premier League. International Journal of Forecasting, 35(2), pp.741-755.

7) Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19, 544–553.
8) Owramipur, F., Eskandarian, P. and Mozneb, F.S., 2013. Football result prediction with a Bayesian network in the Spanish League, the Barcelona team. International Journal of Computer Theory and Engineering, 5(5), p.812.
9) Godin, F., Zuallaert, J., Vandersmissen, B., De Neve, W. and Van de Walle, R., 2014, June. Beating the bookmakers: leveraging statistics and Twitter microposts for predicting soccer results. In KDD Workshop on large-scale sports analytics (pp. 2-14). New York, NY, USA: ACM.
10) Kampakis, S. and Adamides, A., 2014. Using Twitter to predict football outcomes. arXiv preprint arXiv:1411.1243.
11) Rathan, M., DeepthiRaj, N., SankethAnupriya, S., & Vishnu (2018). Football match outcome prediction using sentiment analysis of Twitter data.
12) International Journal of Advanced Research in Computer Science, 9, 78-80.Kabakus, A.T., Simsek, M. and Belenli, Y., 2018. The wisdom of the silent crowd: predicting the match results of World Cup 2018 through Twitter. International Journal of Computer Applications, 182, pp.40-45.
13) Miranda-Peña, C., Ceballos, H.G., Hervert-Escobar, L. and Gonzalez-Mendoza, M., 2021, June. Predicting soccer results through sentiment analysis: A graph theory approach. In International Conference on Computational Science (pp. 422-435). Cham: Springer International Publishing.
14) Le, L., Ferrara, E. and Flammini, A., 2015, November. On predictability of rare events leveraging social media: a machine learning perspective. In Proceedings of the 2015 ACM Conference on Online Social Networks (pp. 3-13).
15) DataReportal. (n.d.).Digital 2025: Global overview report. DataReportal. https://datareportal.com/?utm_source=DataReportal&utm_medium=Social_Platform_Page&utm_campaign=Digital_2025&utm_content=Social_Platform_Page_Home_Page_Link [Accessed 18 July 2025]
16) Alhadlaq, A. and Alnuaim, A., 2023. A Twitter-based comparative analysis of emotions and sentiments of Arab and Hispanic football fans. Applied Sciences, 13(11), p.6729.