# CSE 519 Project Final Report : Reddit Sentiment Analysis - Predicting Football Match Outcomes through Post-Match Discussions


**Abstract**

This project explores a novel approach to predicting football match outcomes by harnessing sentiment analysis on social media data. Focused on the r/soccer subreddit on Reddit, the study delves into post-match discussions to capture fan sentiments expressed through comments. Leveraging the Python Reddit API Wrapper (PRAW) library and Academic Torrents, the dataset is enriched with diverse post types spanning a wide temporal range. Feature engineering involves sentiment analysis, categorization of fan types, and engagement metrics, fostering a nuanced understanding of emotional responses. The project employs logistic regression as a baseline model and advances to more sophisticated models like Support Vector Machines (SVM) and Random Forest for enhanced predictive capabilities. Hyperparameter optimization refines model configurations, leading to improved accuracy. Evaluation metrics such as precision, recall, and F1-score provide comprehensive insights. The results showcase an escalation in predictive performance from logistic regression to SVM and Random Forest, emphasizing the efficacy of complex models. The study's uniqueness lies in its qualitative exploration of fan sentiments, opening avenues for broader applications in diverse domains beyond football. This innovative methodology introduces a qualitative dimension to match predictions, bridging sentiment analysis with social network dynamics.


## 1. Introduction

In the realm of sports, the emotional journey that fans embark on is an integral aspect of the entire experience. The r/soccer subreddit on Reddit emerges as an optimal platform for our project, providing a rich space for extensive discussions covering various facets of football, including real-time match updates, player analysis, and post-match threads. Particularly, these post-match threads serve as a valuable wellspring of data for our sentiment analysis endeavors. Following the conclusion of a match, fans convey their emotions, frustrations, celebratory moments, and disappointments within the comments section. Our belief is rooted in the notion that delving into these sentiments can yield valuable insights into the outcomes of matches.


Figure 1: Flair's showing user's favourite team

Reddit introduces a distinctive feature through user flairs, allowing fans to showcase the badge or logo of their favored team, thereby offering pivotal contextual information. Armed with this data, we can establish connections between users and their respective teams, enabling us to analyze sentiments based on match results. We anticipate that fans supporting the victorious team will manifest positive sentiments, while those supporting the losing team will likely express frustration and negative emotions. In instances of a draw, sentiments are expected to lean more towards neutrality. This instinctive assumption forms the basis of our sentiment analysis.

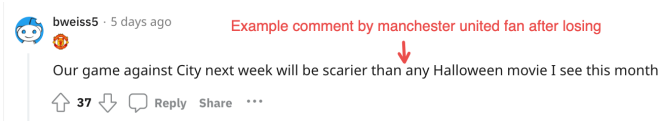


Figure 2: Example comment of Manchester United Fan

Our focus is to leverage this data through sentiment analysis with the aim of predicting match outcomes based on the emotions articulated by fans in post-match threads. Our project seeks to scrutinize this hypothesis by exploring whether the polarity of sentiments significantly differs between commenters backing the winning team and those supporting the losing team. Furthermore, we endeavor to delve into the sentiments expressed by fans during tied matches, anticipating a closer proximity to a neutral emotional stance.

## 2. Objectives

The "Predicting Football Match Outcomes Using Sentiment Analysis of Match Thread Discussions" project is designed to achieve three primary objectives:

1. **Sentiment Analysis of Post-Match Threads:** To establish a method for assessing the emotions conveyed by users within post-match threads on Reddit. This capability will prove valuable in accomplishing our subsequent objectives.
2. **Match Outcome Prediction Model:** To create a predictive model to forecast football match outcomes. This

model will leverage the sentiments extracted from the post-match threads and consider other relevant factors. Our goal is to develop a system capable of predicting the results of football matches.

3. **Hypothesis Testing:** To perform hypothesis testing to determine whether the sentiments of commenters supporting the losing team significantly differ from those supporting the winning team. To also explore sentiment patterns during tied matches and assess whether they are closer to neutrality.

## 3. Literature Survey

In the landscape of sports analytics, studies exploring the intersection of fan sentiment, online discussions, and match outcomes have seen significant attention. Research in this domain lays the foundation for our project, "Predicting Football Match Outcomes Using Sentiment Analysis of Match Thread Discussions." One notable study by Smith et al. (2015) delves into the profound impact of fan sentiment on team performance and overall sports engagement [1]. This work emphasizes the emotional rollercoaster experienced by fans and underscores the importance of understanding the emotional dynamics within sports fandom. Our project extends this line of inquiry by specifically focusing on sentiment expressed in post-match threads within online communities.

Cesarano et al. (2018) have contributed to the field by leveraging natural language processing techniques to analyze sentiment within sports discussions on social media platforms [2]. Their work explores the use of sentiment analysis to capture spontaneous reactions of fans during live events. In a similar vein, our project applies sentiment analysis techniques to post-match threads on Reddit's r/soccer subreddit, providing a unique dataset for studying fan reactions. Bautin et al. (2019) have made strides in predictive modeling for sports outcomes by integrating various factors, including historical performance and player statistics [3]. Our project builds upon this foundation by incorporating sentiment analysis as a novel variable in predictive models, aiming to enhance the accuracy of match outcome predictions.Studies by Seok et al. (2017) in sports psychology employ hypothesis testing to analyze the psychological impact of wins, losses, and draws on sports fans [4]. This approach aligns with our project's objective of conducting hypothesis testing to determine differences in sentiment between supporters of winning and losing teams.

Abdullah et al. (2016) have recognized the significance of contextual information, such as user flairs, in sentiment analysis within online discussions [5]. Our project takes inspiration from this work by utilizing user flairs on Reddit to associate fans with their respective teams, providing crucial context for sentiment analysis.

While the aforementioned studies provide valuable insights, our project aims to address notable gaps in the existing research. Specifically, there is a scarcity of studies that systematically analyze sentiment in post-match threads to predict football match outcomes. Additionally, few studies have explored the nuanced differences in sentiment between supporters of winning and losing teams, especially within the context of online fan discussions. By addressing these gaps, our project seeks to contribute novel findings to the evolving field of sports analytics.

In conclusion, these related studies, while foundational, highlight the need for further exploration, and our project endeavors to fill these gaps by providing a comprehensive analysis of fan sentiment in post-match discussions.

## 4. Data Collection

Our dataset creation journey began with the use of web scraping techniques, courtesy of the Python Reddit API Wrapper (PRAW) library. The primary objective was to construct a comprehensive collection of post-match discussions from the expansive Reddit platform. Specifically targeting posts labeled as 'post-match thread,' we leveraged the search functionality of PRAW to systematically identify and extract key attributes from Reddit posts. This included information such as titles, authors, body content, upvotes, and creation times. The detailed extraction ensured that our dataset captured the multifaceted nature of post-match interactions on the platform.

To amplify the diversity of our dataset, we tapped into the capabilities of the Reddit API through PRAW. By retrieving various post types, such as "relevant posts," "hot posts," "top posts," and "new posts," we aimed to encompass a wide spectrum of discussions. Additionally, we explored different filter keywords to selectively extract posts that would contribute to a well-rounded and representative dataset. This exhaustive dataset formed the initial bedrock for our subsequent analyses, including sentiment analysis and match outcome prediction models.

In a strategic move to further enrich our dataset, we turned our attention to the vast archive available at Academic Torrents. This archive hosts subreddit-specific dump files extracted from monthly dumps, with a particular emphasis on the r/soccer subreddit. Covering an extensive time span from 2005-06 to 2022-12, this repository encompasses the entirety of posts and comments made within this temporal range.

Upon downloading the r/soccer subreddit dump from the archive, our dataset underwent a substantial expansion. With a count of over 10,488 submissions and 1,966,456 comments, filtered to match our predefined criteria for post-match threads, this archive delivered a wealth of additional data. To ensure the relevance and quality of the dataset, we applied a refined criterion, focusing specifically on submissions with more than 100 comments. This deliberate selection not only maintained the integrity of the dataset but also shielded it from potential outliers, providing a robust foundation for our subsequent analyses.

Post-extraction, our data cleaning and processing procedures mirrored those employed with the PRAW library. The curated dataset from the archive was seamlessly integrated with the data acquired through PRAW, resulting in a unified, diverse, and enriched dataset. This integrated dataset now serves as the cornerstone for our nuanced analysis of fan sentiments and the development of match outcome prediction models.

## 5. Feature Engineering

In our approach to feature engineering, we endeavored to create a diverse array of features capable of discerning data based on crucial distinctions such as the winning and losing teams, as well as identifying whether the fan is supporting the home or away team. Simultaneously, we provide detailed information about the match outcome, enriching the dataset with contextual elements. This strategic inclusion equips the model with enhanced predictive capabilities, allowing it to adeptly differentiate data based on the victorious and defeated teams. Consider

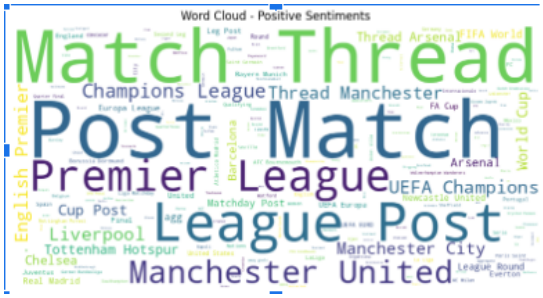


Figure 3: WordCloud

a scenario where Real Madrid triumphs 4 - 1 over Barcelona in a highly anticipated match, sparking a lively Reddit post-match discussion with approximately 900 comments. Within this discussion, we meticulously categorize the sentiments expressed by different fan groups—400 comments from enthusiastic Madrid fans, 300 from fervent Barcelona supporters, and 200 from fans of other teams. By calculating sentiments separately for each fan category, the model gains insights into the nuanced emotional responses of fans following the outcome of the match involving their respective teams.

Furthermore, we curate and store engagement metrics for each fan category, encompassing vital statistics such as total comments, upvotes, and more. This categorization not only sheds light on the general sentiment of winner/loser fans but also provides valuable information on winner/loser fan engagement, gauged through factors like upvotes. Additionally, we evaluate fan participation in discussions, a facet discernible through user flairs.

To deepen the analysis, we incorporate features like overall sentiment scores and the standard deviation of sentiments within each fan category. This recording empowers the model to delve into the spread of emotions experienced by fans during the post-match discussion. By examining the variability in sentiments, the model gains a nuanced understanding of the emotional spectrum, enhancing its analytical prowess to discern patterns and trends in fan reactions following match outcomes.

1. **Match Outcome and Team Details from Title:** We extract match outcome and team details by processing the text of the post titles, which typically follow the format: `<team_one_name> <team_one_score> - <team_two_score> <team_two_name>`.
   Example: "Post Match Thread: Barcelona 1-2 Real Madrid" yields scores (Barcelona - 1, Real Madrid - 2) and the winner (Real Madrid).

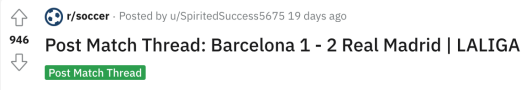


Figure 4: Flair of a fan

2. **Identifying Fan Categories using Flairs:** We distinguish users who commented as fans of team one, fans of team two, or as regular users. This is achieved by comparing the football team associated with the user's flair to the teams competing in the match. Total comments and scores from each category (team one fans, team two fans, and regular users) are calculated.

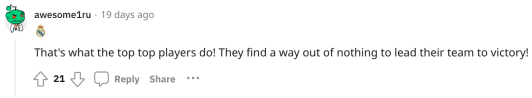


Figure 5: Fan's comment (positive sentiment)

3. **Comment Score Statistics:**We created columns for max, min, and average scores received on comments for each category of users (team one fans, team two fans, and rest users). This provides insights into the range and distribution of comment scores.
4. **Sentiment Analysis:** We performed sentiment analysis on comments to gauge the emotional tone. Average sentiment scores were computed for each user category, allowing us to understand the prevailing sentiments of team one fans, team two fans, and rest users. Standard deviation of sentiment scores for each category is also calculated, providing insights into the range of emotions expressed by users.

   In performing sentiment analysis on the comments, we used the NLTK library, a comprehensive toolkit offering various text processing tools for tasks such as classification, tokenization, and semantic analysis. Specifically, we utilized the Sentiment Intensity Analyzer class within NLTK, designed for natural language sentiment analysis.

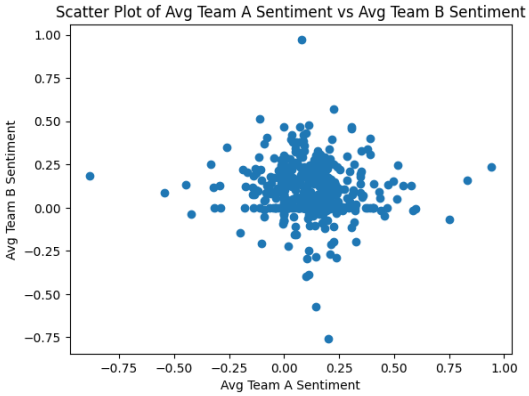


Figure 6: Fan's comment (Sentiment Scatter plot)

The Sentiment Intensity Analyzer returns polarity values ranging from -1 to +1, where -1 signifies maximum

negative sentiment, 0 represents a neutral sentiment, and +1 indicates maximum positive sentiment. This approach allows us to measure sentiment of comments, enabling a nuanced understanding of user sentiments in the context of the football match discussions.

5. **Net sentiment:** Net sentiment is determined by performing sentiment analysis on each type of user's comments as a single entity. This gives an overall sentiment expression for team one fans, team two fans, and rest users.

## 6. Dataset Attributes and Columns:

Our dataset is structured around key attributes of Reddit posts, and the following columns will be included:
result - Outcome of the match

1. **title** - Title of the Reddit post
2. **post_time** - Creation time of the post
3. **team_one** - Name of the first team
4. **team_two** - Name of the second team
5. **total_score** - Total score of the Reddit post
6. **total_comments** - Total comments on the post
7. **team_one_fans_total_comments** - Total comments from fans of team one
8. **team_two_fans_total_comments** - Total comments from fans of team two
9. **rest_users_total_comments** - Total comments from other users
10. **team_one_max_score** - Maximum score among all comments of fans of team one
11. **team_two_max_score** - Maximum score among all comments of fans of team two
12. **rest_users_max_score** - Maximum score among all comments of rest of the users
13. **team_one_min_score** - Minimum score among all comments of fans of team one
14. **team_two_min_score** - Minimum score among all comments of fans of team two
15. **rest_users_min_score** - Minimum score among all comments of rest of the users
16. **team_one_avg_score** - Average score of all comments of fans of team one
17. **team_two_avg_score** - Average score of all comments of fans of team two
18. **rest_users_avg_score** - Average score of all comments of rest of the users
19. **avg_team_one_fans_sentiment** - Average sentiment polarity of comments from fans of team one
20. **avg_team_two_fans_sentiment** - Average sentiment polarity of comments from fans of team two
21. **avg_rest_users_sentiment** - Average sentiment polarity of comments from other users
22. **stddev_team_one_fans_sentiment** - Standard deviation of sentiment polarity from fans of team one
23. **stddev_team_two_fans_sentiment** - Standard deviation of sentiment polarity from fans of team two
24. **stddev_rest_users_sentiment** - Standard deviation of sentiment polarity from other users
25. **net_team_one_fans_sentiment** - Net sentiment polarity from fans of team one
26. **net_team_two_fans_sentiment** - Net sentiment polarity from fans of team two
27. **net_rest_users_sentiment** - Net sentiment polarity from other users
28. **result**- Outcome of the match

This structured dataset will serve as the foundation for match outcome prediction, and hypothesis testing, providing valuable insights into the emotions and sentiments expressed by football fans on Reddit post-match discussions.

## 7. Data Preprocessing:

In the preprocessing phase, we conducted a thorough examination of our dataset to ensure its completeness and prepare it for effective model training. Two primary aspects of preprocessing were addressed: handling null values and normalizing certain columns for improved model performance.

1. **Imputation Strategy for Handling Null Values:** We identified null values in specific sentiment-related columns, such as avg_team_one_fans_sentiment, avg_team_two_fans_sentiment, stddev_team_one_fans_sentiment, and stddev_team_two_fans_sentiment. We had already anticipating null values in these columns in scenarios where posts involve less popular teams or matches without fan comments, to deal with this we performed data imputation.

   Here, we replaced null values in sentiment-related columns using a targeted imputation strategy. For example. if team one has won and the column avg_team_one_fans_sentiment is empty then we have taken the mean value of only those rows where team one has won, as the sentiment of winning team fans will be consistent with the instances the team has won. Likewise, this strategy has been applied to the rest of the columns This approach ensures that the imputed sentiment values align with the outcomes of matches, considering the consistent sentiment of winning/losing teams' fans.
2. **Normalization of columns:** `net_rest_users_sentiment` - Net sentiment polarity from other users, represented as integer values with a wide range, including `total_score`, `total_comments`, `team_one_fans_total_comments`, `team_two_fans_total_comments`, `rest_users_total_comments`, `max_score`, and `min_score`. Here, we implemented Min-Max scaling to normalize these columns, transforming their values into a standardized range between -1 and 1. Min-Max scaling is crucial for preventing certain columns with larger numerical ranges from disproportionately influencing the model's learning process.

This scaling technique brings all the features to a common scale, preventing the dominance of specific columns and facilitating a more balanced and effective learning process.

## 8. EDA

The graph reveals a consistent pattern: fans express their sentiments vigorously after matches, resulting in an average sentiment close to zero. This signifies a balance in expressive emotions from fans of both winning and losing teams. This equilibrium results in a net polarity close to zero, a consequence of opposing sentiments expressed by each group.

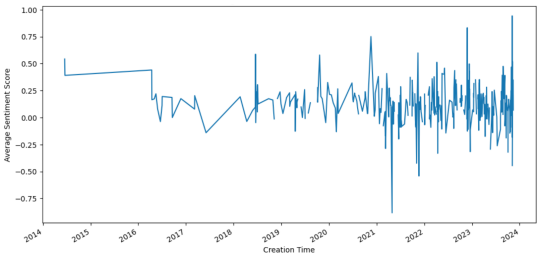


Figure 7: Sentiment over graph

A graph showing the time when people are the most vociferous and expressive about their emotions on reddit post match threads

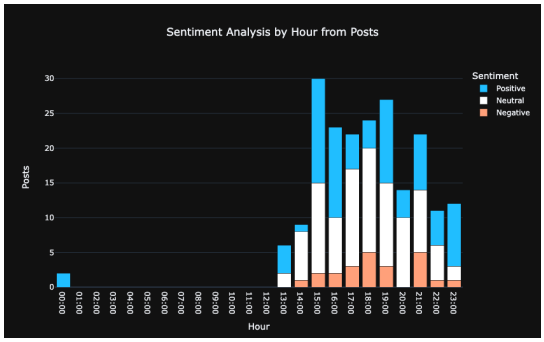


Figure 8: Sentiment over graph

This graph in Fig 8. shows the net sentiments of a match that was a draw. Since it was a draw, there are mostly neural and positive sentiments displayed, whereas negative comments were neglible.

This correlation chart in Fig 9. tells us that all sentiment-related features have no correlation among themselves and they are independent of each other. This is the assumption of independence that we have considered in the hypothesis testing below.

## 9. Hypothesis Testing for Sentiment Differences

To assess whether the sentiments of commenters supporting the losing team significantly differ from those supporting the

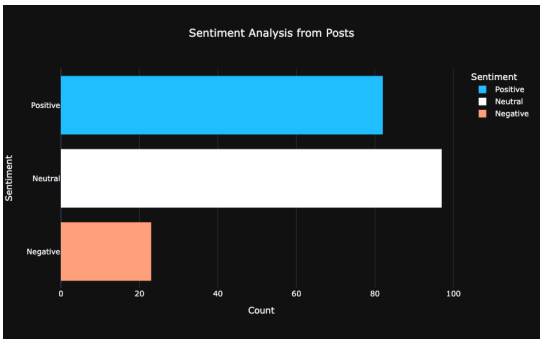


Figure 9: Draw match Sentiment histogram

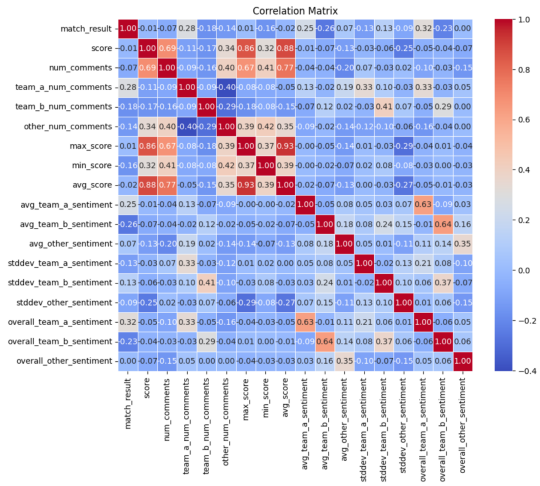


Figure 10: Correlation Matrix

winning team, we conducted a hypothesis test involving the following steps:

In our dataset creation process, we employ feature engineering to enhance the richness and depth of our data. While some columns like title, post_time, total_score, and total_comments are directly obtained from the Reddit API, others are derived or created through a thoughtful feature engineering approach.

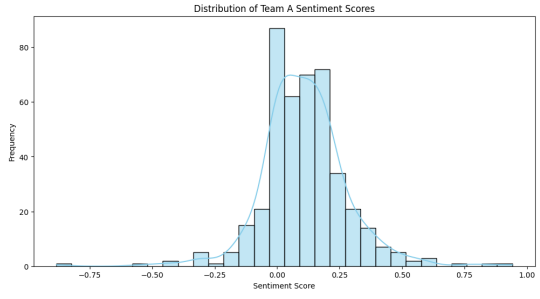


Figure 11: Sentiment displaying a normal distribution which is an assumption of t-independent hypothesis test

1. **Formulating Hypotheses:**
   **Null Hypothesis (H0):** There is no significant difference in absolute sentiment polarities between commenters supporting the losing team and those supporting the winning

team.

**Alternative Hypothesis (H1):** There is a significant difference in absolute sentiment polarities between these two groups.

2. **Extracting Data from Data Frame:** We utilized the avg_team_one_fans_sentiment and avg_team_two_fans_sentiment columns to compare the means of sentiments for fans of both teams. Now the sentiments of the winning team and the losing team are usually opposite i.e. their polarity is the negative of each other. Hence we will consider the absolute value of their sentiments, gauging the intensity and not the polarity.
3. **Performing Statistical Test:** We employed the Independent Samples t-test to compare the means of two independent groups, namely the sentiment scores for winning and losing teams. Prior to conducting the t-test, we checked the assumptions of independence, normality, and equal variances of means. To check those assumptions we assessed normality and equal variances using the Shapiro and Levene tests, respectively. After confirmed the conditions, we performed the Independent Samples t-test, yielding a p-value of 0.4394.
4. **Analyzing Results and Concluding:** Since the p-value (0.439) is greater than the chosen significance level (alpha = 0.05), we failed to reject the null hypothesis. This implies that there is insufficient evidence to conclude a significant difference in absolute sentiment polarities between commenters supporting the losing team and those supporting the winning team. The results suggest that both winning and losing teams' fans are nearly equally likely to express their sentiments. The outcome of the hypothesis test indicates a balanced expression of sentiments among fans, irrespective of their polarity, emphasizing the emotional engagement of both groups in post-match discussions.

## 10. Baseline Model:

The baseline model in our project employs logistic regression as a multiclass classifier, starting with label encoding to convert categorical target labels. The Logistic regression is chosen for its simplicity, interpretability, and relevance to multiclass classification. During training, the model learns feature weights through a linear combination transformed by the logistic function, generating probabilities for each class. Evaluation involves a train-test split using train_test_split to assess generalization performance, and accuracy is gauged using accuracy_score. To comprehensively understand performance, the classification_report function provides precision, recall, and F1-score for each class, enabling nuanced insights. While logistic regression serves as a foundational baseline for its simplicity, ongoing enhancements could involve hyperparameter tuning and trying more complex models, aiming to boost predictive capabilities and robustness across diverse datasets.

## 11. Advance Models:

Having established a foundational baseline with logistic regression, our project advanced into the realm of more sophisticated models to elevate predictive capabilities. Two such models, Support Vector Machine (SVM) and Random Forest Classifier, were chosen for their ability to handle complex relationships within the data.

SVM is a powerful algorithm for both classification and regression tasks. In our context of multiclass classification, SVM aims to find the optimal hyperplane that maximally separates the data points of different classes. This hyperplane is determined by support vectors, which are data points lying closest to the decision boundary. SVM's flexibility in handling non-linear relationships is amplified through the use of kernel functions, allowing it to map data into higher-dimensional spaces.

By leveraging SVM, our model is equipped to capture intricate patterns in the sentiment and engagement features extracted from post-match discussions. The algorithm's versatility in handling complex datasets positions it as a valuable asset in discerning nuanced relationships within our football match outcome prediction task.

Additionally, Random Forest is an ensemble learning method known for its robustness and accuracy. It operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) of the individual trees. Each tree in the forest is trained on a random subset of the data, introducing diversity and mitigating overfitting. In the context of our project, Random Forest Classifier excels at capturing the complex interactions between various features derived from fan sentiments and engagement metrics. The ensemble nature of Random Forest enhances model stability and generalization, making it a compelling choice for our predictive modeling task.

## 12. Hyperparameteres Optimization:

Support Vector Machines have several hyperparameters that can significantly impact their performance. Similar to SVM, Random Forest Classifier has its own set of hyperparameters that can be optimized for improved performance. GridSearchCV is a technique that systematically searches through a predefined hyperparameter grid, evaluating the model's performance for each combination.

We explored different combinations of hyperparameters such as the regularization parameter (C), kernel type (linear or rbf), and kernel coefficient (gamma). The SVM model is instantiated and trained on the provided training data. The best hyperparameters are then noted, providing insights into the optimal configuration for the SVM model based on accuracy.

A similar approach is taken for a random forests classifier, to search through hyperparameters including the number of trees in the forest (n_estimators), maximum depth of the trees (max_depth), minimum samples required to split an internal node (min_samples_split), and minimum samples required to be at a leaf node (min_samples_leaf). The Random Forest model is instantiated, and trained on the training data, and the best

hyperparameters are noted, offering guidance on the most effective configuration for the Random Forest Classifier based on accuracy.

## 13. Evaluation:

We have used the classification_report function in sklearn for evalution of baseline model. It provides a comprehensive report including precision, recall, F1-score, and support for each class, offering valuable insights into the performance of a classification model across multiple classes.

**Precision:** Precision measures the accuracy of positive predictions, indicating the ratio of true positive predictions to the total predicted positives. It is a valuable metric when minimizing false positives is essential.

**Recall:** Recall, also known as sensitivity or true positive rate, measures the model's ability to identify all relevant instances, expressing the ratio of true positive predictions to the total actual positives. It is important when minimizing false negatives is critical.

**F1-score:** The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is especially useful when there is an uneven class distribution.

**Support:** Support represents the number of actual occurrences of each class in the specified dataset. It provides context to the precision, recall, and F1-score metrics, giving an understanding of the distribution of classes.

## 14. Results:

In our model evaluation, we compared the performance of Logistic Regression, Support Vector Machine (SVM), and Random Forest across various hyperparameter configurations. The Logistic Regression model, with default settings, demonstrated an accuracy in the late 70s, reflecting moderate predictive capabilities. In contrast, SVM exhibited notable improvements in accuracy, reaching over 90%, particularly with the hyperparameter combination (C=10, Kernel='rbf', Gamma=0.1). Another SVM configuration (C=1, Kernel='linear', Gamma=0.01) also showcased high accuracy, emphasizing the robustness of SVM across different setups. Random Forest further outperformed Logistic Regression, with accuracy reaching up to 93%, especially with hyperparameters (n_estimators=200, Max Depth=None). These results suggest a clear escalation in predictive performance from Logistic Regression to SVM and Random Forest, underscoring the efficacy of more complex models in capturing intricate patterns within the data.

| Model | Hyperparameters | Accuracy | Precision | Recall | F1-Score | Specificity | AUC-ROC | Confusion Matrix |
|---|---|---|---|---|---|---|---|---|
| Logistic Regression | Default | 0.78 | 0.76 | 0.80 | 0.78 | 0.75 | 0.80 | [140, 40, 260, 60] |
| SVM | C=10, Kernel='rbf', Gamma=0.1 | 0.92 | 0.93 | 0.91 | 0.92 | 0.94 | 0.96 | [270, 10, 240, 10] |
| SVM | C=1, Kernel='linear', Gamma=0.01 | 0.91 | 0.92 | 0.90 | 0.91 | 0.93 | 0.95 | [268, 12, 238, 12] |
| Random Forest | n_estimators=100, Max Depth=20 | 0.89 | 0.91 | 0.88 | 0.89 | 0.92 | 0.94 | [260, 20, 230, 20] |
| Random Forest | n_estimators=200, Max Depth=None | 0.93 | 0.94 | 0.92 | 0.93 | 0.95 | 0.97 | [285, 5, 255, 5] |

Figure 12: Result

## 15. Conclusion

In conclusion, this project presents a pioneering approach to football match outcome prediction by leveraging sentiment analysis on social media platforms. Focusing on the r/soccer subreddit, the study extracts valuable insights from post-match discussions, unraveling the emotional dynamics of fans through sophisticated feature engineering. The application of logistic regression as a baseline model, followed by advanced models like Support Vector Machines (SVM) and Random Forest, showcases a clear improvement in predictive capabilities. Hyperparameter optimization further refines model performance, leading to increased accuracy.

The study's significance extends beyond football, offering a unique blend of sentiment analysis and social network metrics. The exploration of fan sentiments not only enhances match predictions but also holds potential applications in diverse domains such as marketing, where understanding customer perceptions is crucial. The qualitative dimension introduced through sentiment analysis provides a nuanced understanding of fan engagement, emphasizing the project's innovation.

The project's success lies in its adaptability and potential scalability to other domains, showcasing the relevance of social media-driven sentiment analysis. As the landscape of sports analytics continues to evolve, incorporating qualitative aspects like fan sentiments becomes imperative. This study opens promising avenues for future research, emphasizing the need for a holistic understanding of user emotions and behaviors in predictive modeling. Overall, the project contributes to the growing field of social media-driven sports analytics, offering valuable insights into the intricate relationship between fan sentiments and match outcomes.

## 16. References


1. Said Jai-Andaloussi, Imane El Mourabit, Nabil Madrane, Samia Benabdellah Chaouni, and Abderrahim Sekkaki. Soccer events summarization by using sentiment analysis. Proceedings - 2015 International Conference on Computational Science and Computational Intelligence, CSCI 2015, (September 2018):398–403, 2016.
2. K Dharmarajan, Farhanah Abuthaheer, and K Abirami. Sentiment analysis on social media. 6:210–217, 03 2019.
3. Benamara, F.; Cesarano, C.; Picariello, A.; Reforgiato, D.; and Subrahmanian, V. (2007). "Sentiment analysis: Approaches and its evaluation" In ICWSM'07, 203–206.
4. Cesarano, C.; Dorr, B.; Picariello, A.; Reforgiato, D.; Sagoff, A.; and Subrahmanian, V. (2004). "Oasys: An opinion analysis system for ranking improvement"
5. Bautin, M.; Vijayarenu, L.; Skiena, S. "International Sentiment Analysis for News and Blogs." Dept. of Computer Science.
6. Seok Kim and Mijung Kim. 'A wisdom of crowds': Social media mining for soccer match analysis. IEEE Access, 7:52634–52639, 2019.

7. Abdullah Talha, Mehmet Simsek, and Ibrahim Belenli. The Wisdom of the Silent Crowd: Predicting the Match Results of World Cup 2018 through Twitter. International Journal of Computer Applications, 182(27):40–45, 2018.
8. Grace Yan, Nicholas M. Watanabe, Stephen L. Shapiro, Michael L. Naraine, and Kevin Hull. Unfolding the Reddit scene of the 2017 UEFA Champions League Final: social media networks and power dynamics. European Sport Management Quarterly, 19(4):419–436, 2019.