Instituto Tecnológico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences

Master of Science in Computer Science

Thesis proposal

Analyzing Fan Avidity for Soccer Prediction

by

Ana Clarissa Miranda Peña

ID

Monterrey, Nuevo León, June, 2020

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members, hereby, recommend that the proposal presented by Ana Clarissa Miranda Peña to be accepted to develop the thesis project as a partial requirement for the degree of **Master of Science in Computer Science**.

Dr. Miguel González M.
Tecnológico de Monterrey
Principal Advisor

Dr. Laura Hervert E.
Tecnológico de Monterrey
Co-Advisor

First Committee Member's name
First Committee Member's institution
Committee Member

Second Committee Member's name
Second Committee Member's institution
Committee Member

Dr. Hugo Terashima Marín
Director of Program in Computer Science
School of Engineering and Sciences

Monterrey, Nuevo León, June, 2020

Contents

Abstract

Beyond being a sport, soccer has built up communities. Fans showing interest, involvement, passion and loyalty to a particular team, something known as Fan Avidity, have strengthen the sport business market. Social Networks have made incredible easy to identify fans' commitment and expertise. Among the corpus of sport analysis, plenty of posts with a well substantiated opinion on team's performance and reliability are wasted. Based on Graph theory, Social Networks can be seen as a set of interconnected users with a weighted influence on its edges. Evaluating the spread influence from fans' posts retrieved from Twitter could serve as a metric for identifying fans' intensity, if adding sentiment classification, then it is possible to score fans' satisfaction. Previous work attempts to engineer new key performance indicators, or apply ML techniques for identifying the best existing indicators, however, there is limited research on sentiment analysis. In order to achieve the Master's Degree in Computer Science, this thesis aims to strengthen a Machine Learning Model that applies polarity and sentiment analysis on tweets, as well as, discovering factors thought to be relevant on a soccer match. The end goal is to achieve a flexible mechanism, which automatizes the process of gathering data before a match and evaluate its sentiment along with historical factors. This way the model will accomplish independence from the type of tournament, league or even sport.

1 Introduction

Soccer is the world's most popular sport in the world, this premise was stated after a study conducted by Bloomberg during FIFA World Cup 2018, where more than 40% of the interviewees considered themselves soccer fans, and in the same year FIFA reported a 4.64 billion dollars revenue[6]. Soccer has proved to be a fast growing market. Comparing the revenue of FIFA World Cup 2018 to Brazil 2014, when revenue was about 2.1 billion dollars, it is clear that FIFA has doubled its earnings between World Cups. However, when looking at fan engagement, it makes soccer an industry of at least a 22.8 billion dollars industry. According to the UEFA Benchmarking Report, in early 2019, European clubs showed a positive cumulative profit of 1.58 billion dollars when adding sponsorship, broadcasting, gate receipts, etc[14].

Like other sports, soccer is unpredictable and constantly changing over time. Many factors inside and outside the field can influence the results of the next match or even an entire season. Finding a model that could predict the end of a game is an exciting task to accomplish. Besides, the sports sector has been hoarding the gambling industry that represents around 30 - 40% of its operations. The size of this industry, only taking into account the online market, is valued at 450 billion dollars[5]. In 2015 bet365, a popular sports betting company, wagered 52.13 billion dollars. Gambling payout gets updated by calculating a probability based on odds' wagers, this probability accounts for a fixed profit margin established by the bet company. Developing a Machine Learning Model that could generate a probability and adjust payout is also a great area of opportunity.

Several statistics can be retrieved during a match, the most common ones are goals, shots, shots on target, possession in percentage, offsides, corners, pass accuracy, shot accuracy, etc. Those variables could lead to non representative conclusions, for example the relation between possession % and pass accuracy. Also, team's historicity is not taken in consideration. Obtaining a set of identical events in soccer is really hard to achieve, two teams may play against each other every six months, but, players and staff could have changed, just like the local stadium, attendance, etc.

In 2019, Barcelona FC was summing up 52.39 million Twitter followers, while Real Madrid had 50.77 millions and Manchester United 20.81 millions just in this Social Network[12]. Even though, those users may not assume the same level of engagement, there is a valuable amount of information that is not being considered for soccer analysis. Attempting to subtract objectivity from the corpus

generated by fans, and scoring those results based on its level of engagement will provide a better understanding on the seasonal performance of the team, also, and it will also create a mechanism that evaluates the level of reliability from each of the opinions. Constructing a general model that outputs sentiment, polarity and an extra feature like level of engagement from text, could be a relevant application on the communication and advertising field.

As stated earlier, it is possible to obtain a considerable of amount of statistics during a match, and retrieve information from Social Networks of fans opinions. However, there is not yet a model that could perform both tasks, while analyzing sentiment too. The solution proposed incorporates two intermediate models and a final model, which makes the match prediction. The first model is related to feature engineering based on statistics, the second model is related to the text mining approach. API-Football will provide match information limited to the tournament results on five European leagues: Spain, England, Germany, Italy and France on a lapse of seven years ago. Using Twitter API, it is expected to gather tweets to analyze their polarity and sentiment and evaluate the reliability using metrics as eigenvectors and other variables that could be achieved from the platform. After obtaining the best key performance indicators and text metrics, the final step is to benchmark multiple Machine Learning models and test them against future soccer games.

Finally, with the three models working independently, scoring a team status may help to predict a match result. One objective is to establish a Fan avidity model, which relates social media impact to sentiment and engagement based on tweets, this will work to provide insights before a match and could also be a disruptive factor on advertising and marketing fields. The main goal is to obtain a mechanism that automates data gathering, assesses common statistics on the match, and is used for predicting any given team, league or sport.

The thesis proposal is organized in the following way: section 2 defines the problem expected to be solved in more detail; section 3 establishes the Hypothesis and research questions being tested; section 4 specifies the objectives during this research; section 5 summarizes basic concepts needed to perform the research; section 6 provides the Methodology in order of the objectives previously exposed, and finally section 7 presents a year and a half work plan to complete this research.

2 Problem definition

Herold[7] proposes a review that collects several studies on the application of Machine Learning in football for improving attacking play. This summary is oriented towards tracking data, over time consuming game events, and aims to encourage machine learning for improving tactical knowledge and performance. Some of the limitations that are found relates to subjectivity, a finite amount of passes and their trajectory do not provide guidance on the quality, it is either good or bad. Herold[7] also emphasizes the observed divergence that is not captured by those methodologies, like psychological factors and contextual factors, and states that research must incorporate some other areas of study, such as team adaptability and communication. Reading through today's literature, soccer research could be classify according to the way it approaches the problem, either it engineers Data Science or it applies Machine Learning.

It is possible to state that in a five year period, a specific team has played against any other team in the league at least ten times. Ten records are not enough information to predict a fair result, and during those five years changes on the team, staff, management and many other factors could have happened. Machine Learning and Soccer tries to figure out the best existing features to build models that could predict results before a game. However, it is an expensive task, where precision relates severely on the amount and quality of the input features, specifically finding repeatable events

in soccer, as well, modeling seasonal team performance is relevant to the problem.

Some Machine Learning models, for example Neural Networks is a black box when feeding soccer statistics to a predictive model. Authors like Chen[3], who applied a Convolutional Neural Network for predicting a game results having as feature the player ability indexes, prove the gap between the computer science field and soccer experts. The goal is to understand the phenomena and make inferences from a technical point of view.

When trying to use a Soccer ML Model into other sports it may be bespoke, making it hard to fit other disciplines. Building a mechanism oriented towards common patterns could be a better generalization for addressing different sports. Chassy[2] proposed the concept of self-organisation from a psychology perspective, this is how dynamics at the local level determines cohesion and coordination at the system level. The author defines team performance as how well the organization is working in the team. The data science approach converge in a formula for teams performance, this formula states that making frequently and accurate passes will acquire possession to generate shooting opportunities, the relationship between passing density and precision had a $r^2 of 0.99$.

Prediction learning model for soccer matches[8] is also a data science study that analyzes 200,000 results from 52 leagues for training two models, the first a Bayesian Model based on the ranking of each team and the second one uses historicity between teams in dispute. In order to prevent the ranking model for penalizing new teams, the score of veterans teams is calculated from a ratio of 80% points from the current season and 20% on the previous season. Feature engineering is key in this research, once the score is calculated, each team is assigned a rank value according to its league. To keep track of historicity, there is a probability value related to win, loss, and draw rates between the teams in the match.

The use of Machine learning can be a powerful tool for discovering variables relevance, however it requires large sets of data, and some of the algorithms provide insufficient guidance on how the model is learning. When applying Machine learning with Data science, it is possible to get better insight on the problem due to the feature transformations it performs, here the limitations are related to overfiting the behavior of the data, and punisihing generalization in other areas of study.

Some research studies, as the one developed by Shields, [21] evaluate influence of users, represented as nodes, to other entities under the Social Network Analysis, this is performed by calculating a value for each an eigenvector by scoring the weight and the importance of the nodes it is connected to. This paper also adds betweenness centrality, which obtains the shortest paths and find the most repetitive nodes, so that the most influential elements in the network are identified. During World Cup 2018, Kababus [10] constructed a database containing 38,371,358 tweets and 7,876,519 unique users, 9 different machine learning models were trained with the 48 matches on the group phase, and tested to predict round 16, and so on. The features considered for this model are detailed information about the user (number of followers, location, likes count, tweets counts, etc) and the tweet (is it a retweet, reply to a user, retweet count, like count, etc), the highest accuracy obtained was 81.25% when using a Multilayer Perceptron algorithm with 30,000 epochs.

Sajjad[18] uses the SentiWordNet for acquiring corpus-based and context-based representations for sentiment analysis, where word classification is considered either positive, negative or objective, into a one-dimensional vector. This study compares text feature transformations as unigrams and bigrams by clustering its semantic and statistical similarity, later on, features are ranked, and those with the lowest scores are eliminated. The author concludes that SVM algorithm obtains better results on the reduced dimensional vector, but when keeping diversity of features Naive Bayes model has a better performance.

As stated earlier, obtaining a good performance on Machine learning models requires huge amounts of data of enough quality in order to learn patterns from it. From a computer science point of

view, data scientists lack of insight on soccer factors, and what is learned from soccer statistics may varies on different leagues, tournaments and obviously other sports. Extracting relevant opinions from soccer experts in Social Networks by applying Graph Theory could clarify model outcomes, as well, reinforce behavioral prediction techniques through sentiment analysis, may consolidate independence on sport or field of study.

3 Hypothesis and Research Questions

A model that reinforces sports statistics by considering factors outside the field as; sentiment, polarity and fan engagement, abstracting Social Networks as graphs, will predict most accurate results previous a match.

- Is Fan avidity a determinant feature when predicting results previous a match?
- What percentage of successful predictions are obtained, when considering fan avidity as an score of factors as sentiment, polarity and fan engagement?
- How is the percentage of correctness compared to a model without considering Fan avidity?
- How is the percentage of correctness on sentiment classification, when applying betweeness evaluation for keeping the text corpus found to be relevant?
- Could social networks' factors describe the current performance of a team, when compared with key indicators in a set of time?
- Based on Graph Theory, Can a user's sentiment be propagated in a whole eigenvector, this means to influence nearest users' sentiments with the same label as the activation user?

4 Objectives

The general objective of this work is to improve the accuracy on soccer predictions before a match. This will be accomplished by benchmarking several Machine Learning Models taking into consideration highly correlated key performance indicators and Social Networks sentiment analysis as polarity.

- Automate search query gathering from Twitter Standard Search API with a rate of 100 tweets and a window of 15 minutes.
- Automate fixture statistics gathering in API Football with a rate of 5000 requests per day.
- Develop a feature engineering mechanism that weights team's ranking with a constant operation cost; calculating averages and scoring performance.
- Develop a model that fits soccer key indicators for dimensionality reduction with a correlation between 0.8 and 1.0.
- Develop a model that evaluates centrality on Social Network users applying betweenness.
- Develop a model that classifies sentiment and polarity on tweets before a match with an accuracy of around 70-80% on average, according to current research reports.
- Develop a model that predicts a match result using features like fan relevance, sentiment and key soccer indicators with an accuracy above the current research that is between 70-80%.

5 Theoretical framework

5.1 Evaluating centrality in social networks

Riquelme[17] proposes two new centralization measures for evaluating networks. The model graph is compound of labels representing the resistance of the actors to be influenced, and the weight of the edges are the power of influence from one actor to another.

The equation 1 of the **Node activation**.

$$\sum_{j \in F_t(X)} W_{ij} \geq f(i) \tag{1}$$

The activation occurs when the sum of the weight of activated nodes connected to i is greater or equal to i's resistance.

The equation 2 of the **Spread of Influence X**.

$$F(X) = \bigcup_{t=0}^{k} F_t(X) = F_0(X) \cup ... \cup F_k(X) \tag{2}$$

Where t denotes the current spread level of X, and X is an initial activation set.

The first measure considered is called Linear Threshold Centrality, and represents how much an actor i can spread his influence within a network, this by convincing his immediate neighbors.

The equation 3 of the **Linear Threshold Centrality**.

$$LTR(i) = \frac{|F(\{i\} \cup neighbors(i))|}{n} \tag{3}$$

The second measure is Linear Threshold Centralization, this defines how centralized the network is, by finding a k-core which is the maximal subgraph such that every vertex has degree at least k.

The equation 4 of the **Linear Threshold Centralization**.

$$LTC(G) = \frac{|F(C\hat{(}G))|}{n} \tag{4}$$

This relation shows that elements outside the core are easier to be influenced.

Kim[11] proposed a formula to address opportunity based on satisfying fan's requirements. Korean National Football team's comments on the match against Uzbekistan on FIFA World Cup 2018 qualifications where ranked using TF-IDF, which reflects the relevance a word has in the document. After that, a cluster algorithm, such as K-Means, was implemented for topic modeling, once the topic was known, it was assigned a satisfaction value given by the Delphi Method.

The equation 5 of the **Delphi satisfaction expression**.

$$TS_i = \frac{\sum_{j=1}^{j_i} CSij}{J_i} \tag{5}$$

CSi,j satisfaction level of the j-th post in the i-th topic, TSi average satisfaction for the i-th topic and Ji total number of post in the i-th topic.

5.2 Sentiment analysis on social media in the field of sports

Schumaker[19] applies sentiment analysis based on a combination of 8 models using either polarity, such as positive, negative and neutral, and tone such as objective, subjective and neutral. This research has an odds-based approach that gathers an odds-makers match balance sheet on demand of the wagers. Sentiment is calculated by normalizing a specific model data against tweets for a particular club and match.

The equation 6 of the **Normalize polarity**.

$$max(\frac{\sum Tweets|Model_n,Club_1,Match_m}{\sum Club_1,Match_m}\frac{\sum Model_n,Club_2,Match_m}{\sum Club_2,Match_m}) \tag{6}$$

When models tested with negative polarity where higher, they could predict a potential loss, whereas models of positive polarity as a possible win.

In contrast, Dharmarajan[4] applied the Multinomial Naive Bayes Algorithm into two main classifiers. The first one is oriented towards an objective tone, this model is trained with a self-made dataset of well-trusted sources, and the second one is a subjectivity classifier that can either label text as positive or negative. This last one achieved 79,50% accuracy over 32,000 instances, while the first one obtained 77,45% when trained with 86,000 records.

Ljajic[13] proposes a sentiment score by quantifying the logarithmic difference of terms in positive and negative sports comments. Again, sentiment classification is seen as a supervised task which requires creating a domain specific dictionary and assigning a tag as positive or negative for each of the terms. The author proposed the principle of logarithmic proportion TF-IDF as a labeling mechanism.

The equation 7 of the **Polarity compute using TF-IDF**.

$$tfidf_p = (1 + tf_p) * log_{10}\left(\frac{N_p}{N_{t,p}}\right) \tag{7}$$

Where $tfidf_p$ is the polarity of the term in positive comments, tf_p is the term frequency in positive comments, N_p is the number of positive documents and $N_{t,p}$ is the number of positive documents with term t. Same procedure will be followed for negative ratio $tfidf_n$, where the larger term will be set as tag.

A methodology for setting terms as stop words is also concluded on this research, by finding boundaries due to the logarithmic difference of the terms, on the paper boundaries were set when accuracy stopped improving.

The equation 8 of the **Logarithmic difference of term**.

$$DifLog_t = log_{10}\left(\frac{tfidf_p+0.001}{tfidf_n+0.001}\right) \tag{8}$$

Jai-Andaloussi[9] aims to summarize highlights in soccer events by analyzing tweets, for scoring text sentiment they recommend the deep learning method implemented in Stanford NLP which categorizes comments from 0 being very negative to 4 being very positive. However, as the intention is to obtain the most relevant tweets, the moving-threshold burst detection technique is used.

The equation 9 of the **Moving threshold**.

$$MT_i = \alpha * (mean_i + x * std_i) \tag{9}$$

Given l as length of the sliding window at time t_i, $N(l_1)$ to $N(l_i)$ where N is the number of tweets, the mean and standard deviation at time i can be calculated. α is the relaxation parameter, and x is a constant between 1.5 and 2.0. A highlight is defined as $N(l_i) > MT_i$.

5.3 Quantifying performance in soccer

Pappalardo[15] analyzed 6000 games and 10 million events and affirms that team's final position in a competition relates to a typical performance. Based on a multidimensional vector of soccer logs events, two key features could be extracted.

The equation 10 of the **Typical absolute performance**.

$$h_A = \left[\bar{x_1}(A)...\bar{x_n}(A)\right] \tag{10}$$

Results in a vector averaging all features of a given team.

The equation 11 of the **Typical relative performance**.

$$r_A = \left[\bar{\delta_1}(A)...\bar{\delta_n}(A)\right] \tag{11}$$

Results in a vector averaging the difference stated as delta from all features between two teams.

5.3.1 Machine learning in soccer

Tax[20] analyzes some candidate features of interest such as:

- Team was in a lower league the previous year
- Number of matches coached by current coach
- Team hired a new coach during the previous month
- Top-scorer suspended or injured
- Top-assist suspended or injured
- Days since previous match

Barron[1] aims to predict player's trajectory by discriminating 347 variables on an average of 90-minute appearances. Using a Neural Network model with two hidden nodes, the algorithm learned to distinguish a particular player between the three groups on English Football Lower League, League Championship and Premier League. The algorithm obtained an accuracy of 78.8% over 966 outfield players.

A linear model for ranking soccer teams is proposed by Peretti[16] arguing that the computation of the dominant eigenvalue and its corresponding eigenvector result in the final ordering of the tournament. The matrix from which the vectors are calculated is the square matrix of the coefficients a_{ij}. a_{ij} represents a nonnegative number showing the results of the game between team i and team j. The author also synthesized a teams score.

The equation 12 of the **Score for the ith participant**.

$$s_i = \frac{1}{n_i} \sum_{j=1}^{N} a_{ij} r_j \tag{12}$$

Where $r_j$indicates the strength of the team j, and n_i the number of games played by team i.

Current research on sentiment analysis in social networks has accomplished good performance by computing probability based methods such as Naive Bayes or TF-IDF ratio on polarity. Some authors like Schumaker[19], who has built up to eight combinational classifiers by testing between polarity and tone, and others such as Dharmarajan[4], who has highlighted the importance of a made to size dictionary in the topic for differencing between subjectivity and objectivity. However, those findings are missing a weight on the relevance of the social network's author, that is why evaluating centrality insight of the social network could work as an attribute for sentiment classification, as well as a method for selecting the most influential corpus, which attempts to reduce the text needed to classify and removes less relevant data. When quantifying key performance indicators on soccer, some techniques such as averaging statistics are easy calculations, self-explanatory and offer a good summarization on performance, this could be maximized when applying data augmentation methods.

6 Methodology

The next set of steps must be followed, in order to build a Machine Learning Model capable of predicting soccer match outcomes based on key performance indicators and sentiment analysis retrieved from tweets before a game.

1. Initial literature review.

   During this phase the Theoretical Framework is considered to integrate most of the feature engineering procedures on the main models. Also, historical data about soccer matches and previous code are analyzed from the Bayesian approach presented by Hervert[8], mainly understanding the way the ranking is assigned, and how the probability from the Bayesian function is calculated. Subtasks: 1. Analyze Hervert's Bayesian Method and implementation code.

2. API Documentation and data collection.

   This step attempts to explore the API architecture and design to retrieve characteristics that could feed both of the models. After identifying the requests and their corresponding endpoints, crawl techniques should be implemented due to the limited petitions each authentified user can make. The Standard search API will be used to retrieve a maximum of 100 tweets that match a specific query per request, since user authentication is limited to 180 requests in a 15 minutes window. The API-football v.3.5.2 (beta) for testing purposes is allowing up to 5000 requests per day and per identification. Maximizing the amount of information data retrieved from those requests is the most important task. Subtasks: 1.1 Define get methods, 1.2 Perform authentication, 1.3 Crawling script on Twitter's API, 2.1 Define get methods, 2.2 Perform authentication, 2.3 Crawling script on API-football.

3. Data cleaning and preprocessing.

   HTTP 2.0 web protocol manages services communication in a JSON format. The data responses from both APIs must be processed into JSON Objects for building two uniform datasets. Data preprocessing on the Sentiment Dataset is a key factor before the extraction of features. It is important to discard trivial information from enormous graphs seen as abstractions of Social Networks, a threshold under the betweenness centrality measure will help to consider only the core users on the complete network, and augmenting the dataset with an eigenvector will reflect the score of influence of a given user. Once the Sentiment Dataset is reduced according to the condition *score of influence greater or equal than the betweenness threshold*, feature engineering can take place. Subtasks: 1.1 JSON into Sentiment Dataset, 1.2 JSON into Key Performance Indicator Dataset, 2.0 Betweenness Reduction on Sentiment Dataset.

4. Data engineering.

   The Key Performance Indicators Dataset should be transformed to generalize key features. Some of the experiments, done during this step, will be used to define a huge set of match history or several datasets grouped by league. Also, a window size for past seasons should be tested for better results. It is proposed to asses fixtures statistics from API-Football on a specific team and on a given season by its relative performance, this involves the calculation of the average of all previous match statistics per season. Some candidate features to be tested are the count of wins, ties and loses of the last five matches, if the team is in a classification position, if the team is in a descending position, red cards on the previous match, injuries on the previous match, and coach antiquity by number of weeks, also, current quotient will also be included. An

overall score of a team will be calculated as the sum of the average statistics multiplied by its quotient, bigger quotients reflects regularity on a team. The Sentiment Analysis Dataset, after tokenization, is expected to apply the TF-IDF algorithm to find the most relevant words, these words will be assigned a polarity by comparing each one with external research classification models, one suggestion is the use of deep learning Stanford NLP. With this, a specialized dictionary for the subject will be created. Substasks: 1. Key Performance Indicators summarization, 1.1 Average statistics, 1.2 Coach Antiquity, 1.3 Team's score, 1.4 Last five matches results, 1.5 Injuries, 1.6 Red cars, 2.1 TF-IDF tokenization, 2.2 Polarity assignment to dictionary with pre-build classifier.

5. Dimensionality reduction model.

   In this step the objective is to reduce dimensionality without interchanging variance, first attempt will be performed by applying PCA, also known as Principal Component Analysis, which reduces high volume information into the most correlated features. Subtasks: 1. Training with unsupervised PCA Model, 2. Testing iteratively.

6. Sentiment classification model.

   The precomputed dictionary will be of help to set a threshold for classifying a text as positive or negative, this deviation will be calculated with the Logarithmic Difference of a term proposed by Liajic[13]. Once utilizing the Sentiment Analysis Dataset, each of the comments will apply polarity of the terms in positive comments, as well as in negative comments and compare its result within the threshold boundaries to classify its polarity. Subtasks: 1. Training against dictionary built on step 4.2.2, 2. Testing iteratively.

7. Predictive model evaluation and results.

   Here, the final model will compact the information retrieved from PCA iteratively for each of the matches and attach the major polarity from the Sentiment classification model. It is expected to execute a multinomial prediction as win, lose or tie. Training and testing will be performed with data from matches and tweets that had already happened and evaluated with metrics such as accuracy, precision, recall and f1-score. It is expected to lay out this problem as a supervised task. Subtasks: 1. Data engineering augmenting PCA variables with Classified polarity. 2. Training with supervised model. 3. Testing iteratively.

8. Research report.

   The last step is to document all the procedures and the results obtained through the research using tools like LaTex. Subtasks: 1. Write down the report.

7 Work plan

The steps set up in the Methodology are expected to be completed in a two year period, starting from February 2020 and ending in December 2021. However, it is planned to finish with a three month looseness due to report iterations and thesis defense appointments, the expected end date is in early September 2021.

Figure 1 presents a Gantt Diagram showing the activities to carry on in order to reach the objectives of the proposed research.

During the end of the first semester, it is expected to implement GET methods, set client connections from the APIs and start the information retrieval code. Over Summer 2020 the initial scripts

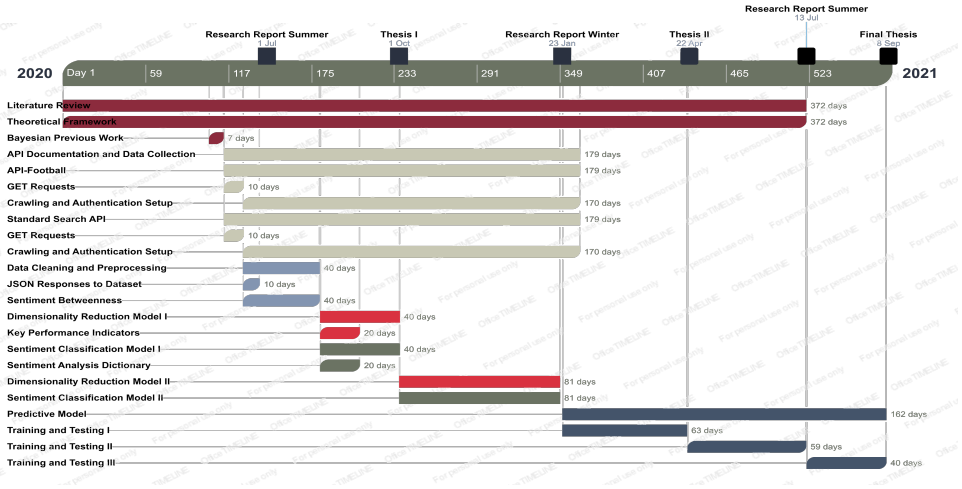

Figure 1: Schedule of activities to complete research project.

formatting the APIs responses into a manageable Dataset must be complete, as well as the data preparation analysis, considering centrality measures on Social Networks. First half of the second semester will be destined to feature engineering on both of the submodels: key performance indicators and sentiment analysis dictionary, as well as an early model stage. The second half is dedicated to training and testing the submodels themselves with the data that was previously transformed. Spring 2021 and Summer 2021 is dedicated to developing the overall model that includes the elective features of the submodels and parameter tuning experiments. Finally, at the end of July 2021 and until early September 2021, the research will be clearly documented.

References

[1] Donald Barron, Graham Ball, Matthew Robins, and Caroline Sunderland. Artificial neural networks and player recruitment in professional soccer. *PLOS ONE*, 13(10):1–11, 10 2018.

[2] Philippe Chassy. Team play in football: How science supports f. c. barcelona's training strategy. *Psychology*, 04:7–12, 01 2013.

[3] Hengzhi Chen. Neural network algorithm in predicting football match outcome based on player ability index. *Advances in Physical Education*, 09:215–222, 01 2019.

[4] K Dharmarajan, Farhanah Abuthaheer, and K Abirami. Sentiment analysis on social media. 6:210–217, 03 2019.

[5] Christina Gough. Sports betting, Mar 2019.

[6] Christina Gough. Soccer, Apr 2020.

[7] Mat Herold, Floris Goes, Stephan Nopp, Pascal Bauer, Chris Thompson, and Tim Meyer. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science Coaching*, 10 2019.

[8] L. Hervert-Escobar, T. I. Matis, and N. Hernandez-Gress. Prediction learning model for soccer matches outcomes. In *2018 Seventeenth Mexican International Conference on Artificial Intelligence (MICAI)*, pages 63–69, Oct 2018.

[9] Said jai andaloussi, Imane Mourabit, Nabil Madrane, Samia Chaouni, and Abderrahim Sekkaki. Soccer events summarization by using sentiment analysis. pages 398–403, 12 2015.

[10] Abdullah Talha Kabakus, Mehmet Şimşek, and Ibrahim Belenli. The wisdom of the silent crowd: Predicting the match results of world cup 2018 through twitter. *International Journal of Computer Applications*, 182:40–45, 11 2018.

[11] Young-Seok Kim and Mijung Kim. "a wisdom of crowds": Social media mining for soccer match analysis. *IEEE Access*, PP:1–1, 04 2019.

[12] David Lange. Largest soccer clubs by size of digital community 2019, Dec 2019.

[13] Adela Ljajić, Ertan Ljajić, Petar Spalević, Branko Arsic, and Darko Vučković. Sentiment analysis of textual comments in field of sport. 09 2015.

[14] Bobby McMahon. Revenue of 22.8b: Uefa report shows the few teams making money and the many that are not, Jan 2019.

[15] Luca Pappalardo and Paolo Cintia. Quantifying the relation between performance and success in soccer. *Advances in Complex Systems*, 21(03n04):1750014, May 2018.

[16] Alberto Peretti. A linear model for ranking soccer teams. *Journal of Interdisciplinary Mathematics*, 22(3):243–263, 2019.

[17] Fabián Riquelme, Pablo González-Cantergiani, Xavier Molinero, and Maria Serna. Centrality measure in social networks based on linear threshold model. *Knowledge-Based Systems*, 140:92–102, 01 2018.

[18] Tofighy Sajjad and Fakhrahmad Seyed Mostafa. A proposed scheme for sentiment analysis: Effective feature reduction based on statistical information of SentiWordNet. *Kybernetes*, 47(5):957–984, jan 2018.

[19] Robert P. Schumaker, A. Tomasz Jarmoszko, and Chester S. Labedz. Predicting wins and spread in the premier league using a sentiment analysis of twitter. *Decision Support Systems*, 88:76 – 84, 2016.

[20] Niek Tax and Yme Joustra. Predicting the dutch football competition using public data: A machine learning approach. 09 2015.

[21] Grace Yan, Nicholas Watanabe, Stephen Shapiro, Michael Naraine, and Kevin Hull. Unfolding the twitter scene of the 2017 uefa champions league final: social media networks and power dynamics. *European Sport Management Quarterly*, pages 1–18, 10 2018.