Predicting Soccer Results through Sentiment Analysis: A Graph Theory Approach

Clarissa Miranda-Peña[1], Hector G. Ceballos, Laura Hervert-Escobar, and Miguel Gonzalez-Mendoza

Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Ave. Eugenio Garza Sada 2501, Monterrey, N.L., Mexico, 64849
A01400214@itesm.mx

Abstract. More than four out of 10 sports fans consider themselves soccer fans, making the game the world's most popular sport. Sports are season based and constantly changing over time, as well, statistics vary according to the sport and league. Understanding sports communities in Social Networks and identifying fan's expertise is a key indicator for soccer prediction. This research proposes a Machine Learning Model using polarity on a dataset of 3,000 tweets taken during the last game week on English Premier League season 19/20. The end goal is to achieve a flexible mechanism, which automatizes the process of gathering the corpus of tweets before a match, and classifies its sentiment to find the probability of a winning game by evaluating the network centrality.

Keywords: Graph Theory · Machine Learning · Sentiment Analysis · Social Networks · Sports Analytics.

1 Introduction

Most of today's literature on Machine Learning and Soccer talks about engineering the best indicators based on match statistics. Also, current research tries to figure out the best existing features to build models that could predict results before a game. However, retrieving data for a set of repeatable events is a difficult task to accomplish, as well, changes on the team, staff, management, and many other factors could have happened. 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 of fans can serve as a metric for identifying fans' intensity. In order to decouple league, team, and even sports information, it is proposed a Sentiment Analysis Model which scores polarity on opinions made by soccer fans on Twitter.

1.1 Review on Social Network Analysis: Spread Influence

Some research studies, as the one developed by Yan, [16] evaluate the influence of users, represented as nodes, on other entities under the Social Network

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

Analysis, this is performed by calculating a value for each 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 finds the most repetitive nodes, so that the most influential elements in the network are identified.

Riquelme[11] proposes two new centralization measures for evaluating networks. The model graph is a compound of labels representing the resistance of the actors to be influenced, and the weight of the edges is the power of influence from one actor to another.

The equation 1 of the **Node activation** is:

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

The activation occurs when the sum of the weight of activated nodes connected to i, in the set of $F_t(X)$ is greater or equal to i's resistance denoted as $f(i)$.

The equation 2 of the **Spread of Influence X** is:

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

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** is:

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

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

The equation 4 of the **Linear Threshold Centralization** is:

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

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

Kim[6] proposed a formula to address opportunity based on satisfying the fan's requirements. Korean National Football team's comments on the match against Uzbekistan on FIFA World Cup 2018 qualifications were ranked using TF-IDF, which reflects the relevance a word has in the document. After that, a clustering 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** is:

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

$CS_{i,j}$ satisfaction level of the *j-th* post in the *i-th* topic, TS_i average satisfaction for the *i-th* topic, and J_i total number of post in the *i-th* topic.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

1.2 Review on Sentiment Analysis

Schumaker[13] applies sentiment analysis based on a combination of 8 models using either polarity, such as positive, negative, and neutral, and tone such as the objective, subjective, and neutral. This research has an odds-based approach that gathers an odds-maker's match balance sheet on demand of the wagers. The sentiment is calculated by normalizing a specific data model against tweets for a particular club and match.

The equation 6 of the **Normalize polarity** is:

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

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

In contrast, Dharmarajan[2] 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[9] 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 that 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** is:

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

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. The same procedure will be followed for negative ratio $tfidf_n$, where the larger term will be set as a 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** is:

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

During World Cup 2018, Talha [14] 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)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

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.

Jai-Andaloussi[4] aims to summarize highlights in soccer events by analyzing tweets and 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** is:

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

Given l as the 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 the 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 $N(l_i) > MT_i$.

2 Methodology

Soccer is constantly changing over time, in order to make this a real-time problem, a framework for gathering recent tweets was built. The methodology is summarized in two key components: first tweets are preprocessed for scoring sentiment polarity, and second, they are evaluated as a Social Network problem by applying graph theory.

2.1 Gathering and preprocessing data

The data is obtained through the Twitter's Standard Search API. The queries were performed in the last match week on Premier League's season 19/20 and were limited to the English language. The data was processed into a Dataframe with the next remaining fields as shown in Table 1. Twitter's documentation on standard operations shows that appending a string happy face ":)" on the query represents a positive attitude, while ":(" represents a negative attitude. The maximum number of tweets retrieved from a request is 100, to aid the evaluation process, three types of queries were performed: first adding the happy face, second adding the sad face, and finally a neutral request without a face.

In the end, a total of 30 JSON requests with a maximum count of 100 tweets were available for study, Table 2 indicates the keywords placed in each fixture query. However, not all fixture requests accomplished these 100 tweets as shown in Table 3.

2.2 Data cleaning

The data cleaning process performed drop of duplicates with a count of one, and drop off empty tweets, the empty tweets were filtered after removing user

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

Table 1. Dataset fields

	Field	Description
1	season	A YYYY representation of the match season.
2	weekgame	The number of the current week match.
3	home_team	A three-letter code abbreviating the home team.
4	away_team	A three-letter code abbreviating the away team.
5	favorite_count	The count of favorites in the tweet.
6	lang	A two-letter code abbreviating the language.
7	retweet_count	The count of retweets in the tweet.
8	retweeted	True or false if the tweet is a retweet.
9	text	The text of the tweet.
10	followers_count	The count of followers from the user.
11	verified	True or false if the account is verified.

Table 2. Queries

	Match	Keywords
1	Arsenal vs Watford	#ARSFC @Arsenal #WatfordFC @WatfordFC #ARSWAT
2	Burnley vs Brighton	#BURFC @BurnleyOfficial #BHAFC @OfficialBHAFC #BURBHA
3	Chelsea vs Wolves	#CFC @ChelseaFC #WWFC @Wolves #CHEWOL
4	Crystal Palace vs Tottenham	#CPFC @CPFC #THFC @SpursOfficial #CRYTOT
5	Everton vs Bournemouth	#EFC @Everton #AFCB @afcbournemouth #EVEBOU
6	Leicester vs Manchester United	#LCFC @LCFC #MUFC @ManUtd #LEIMUN
7	Manchester City vs Norwich	#MCFC @ManCity @NorwichCityFC #MCINOR
8	Newcastle vs Liverpool	#NUFC @NUFC #LFC @LFC #NEWLIV
9	Southampton vs Sheffield Utd	#SaintsFC @SouthamptonFC #SUFC @SheffieldUnited
10	West Ham vs Aston Villa	@WestHam #AVFC @AVFCOfficial #WHUAVL

mentions and ended up with no text to analyze, this returned a count of 11 empty tweets. At first, the library langdetect was used with a threshold of 50% probability for the English language, however, in practice, the language detection

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

Table 3. Requests

	Match	# of *tweets*
1	Arsenalvs Watford	*247*
2	Burnley vs Brighton	*121*
3	Chelsea vs Wolves	*235*
4	Crystal Palace vs Tottenham	*156*
5	Everton vs Bournemouth	*143*
6	Leicester vs Manchester United	*256*
7	Manchester City vs Norwich	*198*
8	Newcastle vs Liverpool	*268*
9	Southampton vs Sheffield Utd	*128*
10	West Ham vs Aston Villa	*175*

accuracy drops drastically on shorter tweets, so multilingual tweets were kept. The length of the dataset finished with a total of 1,915 tweets.

2.3 Data engineering

In order to make the most of the available resources, extra variables were included, two of them relate to text transformations.

- pre_label: integer field that pre classifies the tweet according to the search query, for positive tweets gives 1, negative -1 and 0 if neutral.
- support: integer field that pre represents the support to a given team if it appears on the tweet a mention or hashtag to the home team returns 1 when away team returns -1 and 0 if both appearances happened.
- no_mentions: string field as a version of the tweet without mentions and removing anything that is not plain text.
- with_emojis: string field as a version of the tweet, this sophisticated text transformation keeps mentions, removes links, and uses the emoji library to encode emojis into text, also a regular expression matches happy and sad faces representation and replace happy faces by the word good and sad faces with the word bad.

Classifying polarity was possible by using available resources, such as the Open Source Library Stanza [10], previously named Stanford NLP. Stanza is a language-agnostic processing pipeline that groups together tokenization, lemmatization, part-of-speech tagging, dependency parsing, and named entity recognition. Stanza has a built-in model for Sentiment Analysis [5], this model is trained as a one-layer Convolutional Neural Network using word2vec which are the resulting vectors when applying bag-of-words on 100 billion articles on Google News.

Since it is possible to provide text previously tokenized to Stanza's pipeline, it was preferable to create tokens using NLTK Tweet Tokenizer as it applies

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

regular expressions to maintain mentions. A mention identifies users with the prefix @, while Stanza Tokenizer split those characters underperforming entity recognition.

Figure 1 shows some word clouds comparing the results given on the assumption of the pre_label tag against Stanza's model evaluation.

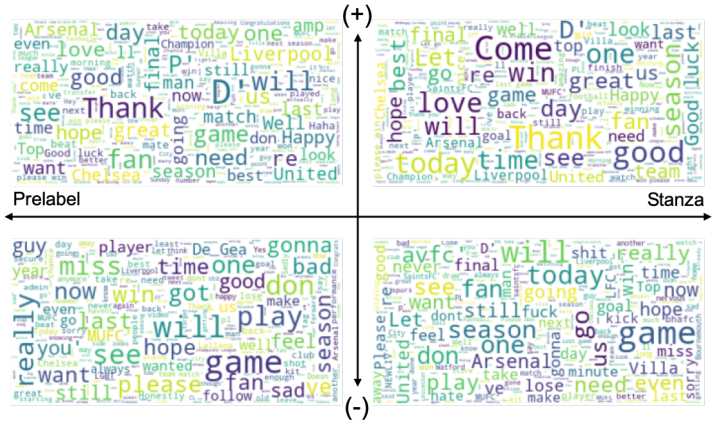

Fig. 1. Word clouds comparison by polarity and model

By applying Stanza's classification it is possible to measure the magnitude of the polarity tags. Now, phrases such as love, good luck, hope, took relevance on the positive tags, while on the negative tags curse words and negations took precedence.

After classifying polarity in tweets from a Machine Learning perspective, two new fields were created a modified support *m_support* and modified sentiment *m_sentiment*. These fields were the result of matching a regular expression that identifies suggested scores of form 0:0 or 0-0 since a result with a goal difference greater than zero indicates clear favoritism to one of the adversarial teams.

Figure 2 shows relevance on the proposed characteristics, this is measured by the amount of neutral support and sentiment that was able to be classified as positive or negative, and as a side of the home team or away team.

2.4 Graph Theory

Popular teams such as Arsenal, Liverpool, Manchester United, etc., have higher rates of tweets, making it difficult to choose a favorite when comparing against its opponent. This section translates the imbalance of favoritism into a graph analysis.

A simple graph [7] has the form $G = (V, E)$ where V is a set of n vertexes and E is a set of n edges. An edge is a link between two vertexes, so an edge E_k is associated with an unordered pair of the vertex (V_i, V_j).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

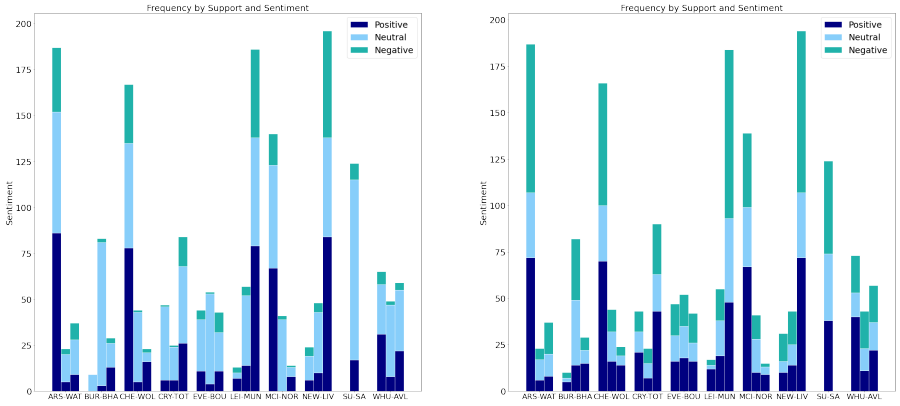

Fig. 2. On the left the frequency of the pre_label polarity by team support, on the right the frequency of the modified polarity by team modified support

Here the vertex are the users and each edge represents a tweet to a tagged user, the tagged user is the team which is mentioned on the tweet. The final tuple looks like: $(fan, team, edge_k)$. Two edges were added to the graph when a tweet mentioned both of the teams. Also, it was preferable to choose a multigraph representation, since it is possible to have multiple edges of a fan to the same team.

Edge's weight Setting a singular value for each edge's weight will miss out on tweets having likes or being retweeted, as well, it would not solve the imbalance problem, since the sum of all edges values from the team's vertex to the fans will be equal to the frequency of the fans of a given team.

The equation 10 of the **Tweet's weight imbalance** is:

$$E_k(U_i,team)=c-\frac{U_i(likes)+U_i(retweets)}{\sum_{i=0}^{k} support(team)+\frac{\sum_{i=0}^{k} support(match|neutral)}{2}} \tag{10}$$

An edge between the user and the team represents the distance the tweet has with the team, whenever a tweet is more retweeted or has a larger amount of likes, it means it is more reachable to the audience of a team, subtracting from a constant c, the relative number of likes and retweets to the support of the team, means reducing the distance between the fan and the team. By calculating the relative influence in a network as the sum of interactions over the frequency of support in a team, a team with fewer followers will represent a greater reach to its network, rather than the reach-in networks with larger amounts of fans, this way the class imbalance problem could be resolved. Neutral support was split in two and added to the frequency of each of the teams as seen in Figure 3 where light blue lines are neutral, navy are the positive tweets and green negatives.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

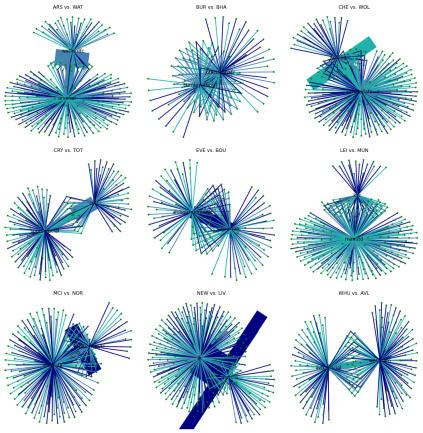

Fig. 3. Graph using tweet's weight imbalance

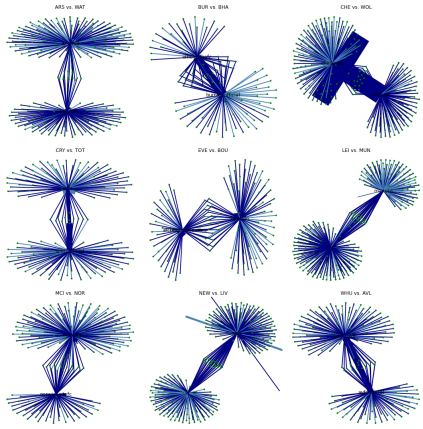

Fig. 4. Graph using tweet's inverted polarity

Inverted polarity This is a counter proposal for solving the imbalance problem, by interchanging support to the adversarial team. This creates a network where negative links to a given team, become positive edges to its opponent and vice versa. Then the network is composed only of positive and neutral polarity represented in Figure 4.

3 Results

3.1 Evaluation

A way for evaluating network entities is through indexing centrality [17], this metric indicates the influence of the vertex in the network. Degree centrality is discarded, since it counts the number of links to a node, and as mentioned earlier there is a clear imbalance between the number of fans, so it might present misleading results. Betweenness centrality is not taken into consideration neither, this measure gives precedence to mediation nodes that connect the network, here users that mentioned both of the teams have the highest scores. Closeness centrality is the selected measure for comparing independence and efficiency of communication in an entity [8].

Closeness centrality It is computed as the reciprocal of the average of shortest-path distance from an agent A_u to all other agents. The equation 11 of the **Closeness centrality** is:

$$C(u) = \frac{n-1}{\sum_{i=1}^{n} d(i,u)} \tag{11}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

Current-flow closeness centrality It is based on information spreading efficiently like an electrical current. Edges are now resistors $r_e = 1/w(e)$ and each vertex has a voltage $v(u)$. The equation 12 of the **Current-flow closeness centrality** is:

$$C(u) = \frac{n}{\sum_{i=1}^{n} v(u) - v(i)} \tag{12}$$

This represents the ratio n to the sum of effective resistances between u and other vertexes quoted [3]. It is also equivalent to the information centrality which considers all path weights, not only the shortest ones, and instead computes its average from the originated vertex. The information in a path is the inverse of the length of a path [1]. The equation 13 of the **Information centrality** is:

$$\bar{I}_u = \frac{n}{\sum_{i=1}^{n} \frac{1}{I_{ui}}} \tag{13}$$

Harmonic centrality Applies harmonic mean to overcome outweighs from infinite distances, and it is computed as the sum of the reciprocal of the shortest path distances. The normalized harmonic centrality can reach up to 1 as the maximum connected vertex. Lower values occur when used on an unconnected graph representing the reduced capability of communication in the network [12]. The equation 14 of the **Harmonic centrality** is:

$$C(u) = \sum_{i=1}^{n} \frac{1}{d(i,u)} \tag{14}$$

3.2 Testing

For evaluation purposes, the study was extended to a set of 54 matches starting at week 38 from season 2019 up to week 8 from current season 2020. In total 7,833 tweets were analyzed. Besides a graph considering the three polarity links, three subgraphs, one for each polarity, were built. Based on centrality measures two cases were considered:

Case 1. Applying current-flow closeness centrality as a comparison measure inter-team, since low resistance will show efficiency in the way a team communicates to its fans. The difference between the current-flow closeness index on the home team and away team will reflect the favorite team given its communication effectiveness. The equation 15 of the **Inter-team closeness** is:

$$diff\_closeness = \|closeness(home) - closeness(away)\| \tag{15}$$

Case 2. Applying harmonic centrality as the leading polarity intra-team, to know which polarity has a better representation of the fan's sentiment towards a team. Communication is more difficult when having fewer connections. For each subgraph, the less fluctuated harmonic centrality given a polarity against the harmonic centrality considering all three polarities will support a good communication capability. The equation 16 of the **Intra-team closeness** is:

$$closeness(\tfrac{team}{polarity}) = \|closeness(team) - closeness(polarity)\| \tag{16}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

Support Vector Machines were used for classifying a match as a win, draw, or lose at home. These models were trained with different centrality indexes and evaluated with five-fold cross-validation. During the pipeline different k best features were applied testing ANOVA.

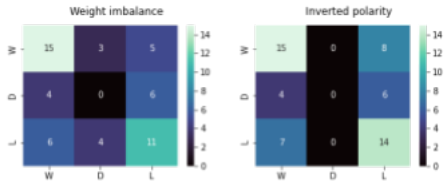

Fig. 5. Confusion matrix comparison

Table 4. Classification report when selecting features from tweet's weight imbalance.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.60	0.65	0.63	23
0	0.00	0.00	0.00	10
1	0.50	0.52	0.51	21
accuracy avg/total	0.48	0.48	0.48	54

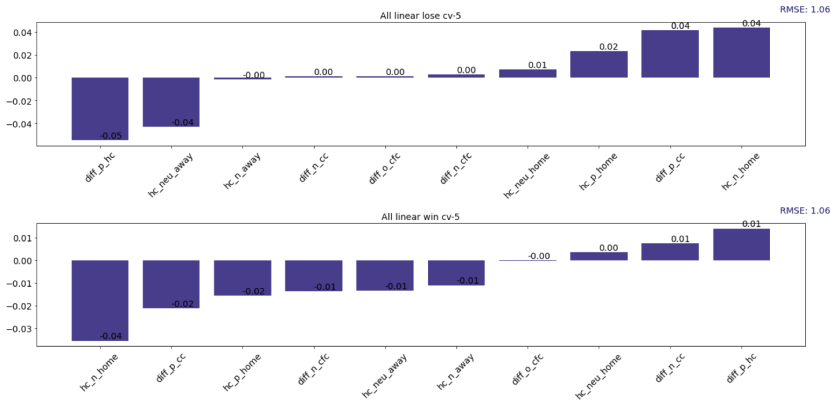

Fig. 6. SVM's selection on tweet's weight imbalance

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

Table 5. Classification report when selecting features from tweet's inverted polarity.

Classification Report				
	precision	**recall**	**f1-score**	**support**
-1	0.58	0.65	0.61	23
0	0.00	0.00	0.00	10
1	0.50	0.67	0.57	21
accuracy avg/total	0.54	0.54	0.54	54

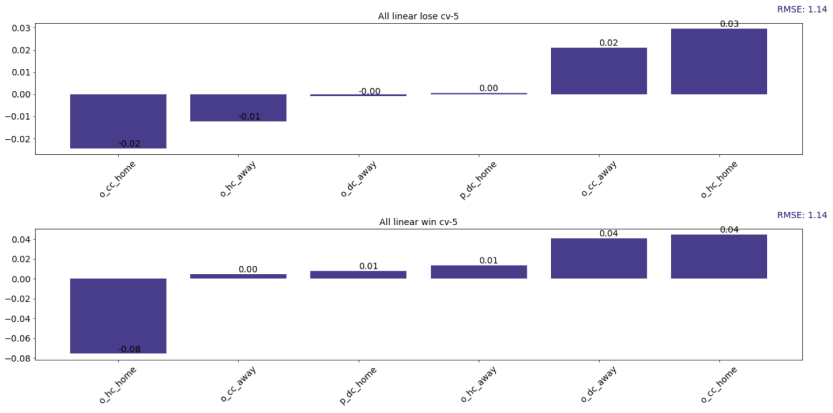

Fig. 7. SVM's selection on the inverted polarity

4 Discussion

Table 5 shows higher values on the recall metric than Table 4, this metric references to all events classified correctly, while the precision metric focus on the predicted positive to be correctly [15]. As seen in Figure 5 although the inverted polarity model has a better recall, the weight imbalance model did not loose precision and gave more diversity to the prediction model by attempting to guess draw matches.

About the metrics used, Figure 6 plots feature weights, and validates the inter-team centrality as the difference on the normalized harmonic closeness centrality, while the intra-team measure is given by singular polarities of a team. Figure 7 confirms the current-flow closeness centrality as a non-significant measure.

Compared to the work done by Schumaker [13] with the highest accuracy of 50.49%, the current model has a slightly better performance with 54% accuracy. However, this is only a sample of all Premier League games, so there is a huge area of opportunity for testing future games.

Finally, this study brought relevant candidate features retrieved from fan expertise in Social Media which can be interchanged with statistics and odds information, for boosting accuracy in soccer prediction before a game.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

5 Conclusion

Football as any other sport is unpredictable, modeling draws is a very difficult task, as well classifying a match previous to its start. Probabilities closer to the end of the game might give more accurate results, but it is also true that a match can be flipped in the last five minutes.

This study gave satisfactory results from the fact that it does not consider statistics at all, instead, it uses as a historical database the knowledge from fans' comments, for scoring polarity to a team and then generates a prediction previous to the start of the game.

This novelty presents a prediction mechanism that can be decoupled from football league's and even sports, whenever comparing a team's sports with scores of the form 0-0, future steps would consider testing this method in other sports. Also as far as our knowledge, it presents a unique methodology for mining Sports sentiment in Social Networks by engineering centrality measures to be considered as candidate features in order to train a Machine Learning model. The proposal of computing edges' weights relative to the size of the network, and to the reach of a tweet by encountering the number of likes and shares, is a unique mechanism for balancing the network. Even, measures as betweenness centrality could lead us to the highest impact fans driven conversation between the networks of the adversarial teams.

Beyond, this study could find a rich area of opportunity in other fields, when scoring polarity tendencies on users by applying intra-team evaluation, and generating comparison metrics between users by an inter-team evaluation. In the marketing sector, the first statement could be seen as the way customers perceive the product, as well, measure an efficient communication to the spectator, while the second could be used as a powerful benchmarking tool.

6 Acknowledgment

The authors are grateful to Tecnologico de Monterrey, who through its Academic Scholarship Program for Graduate Students provided technical and financial support for the development of this research. In addition, we are grateful to CONACyT for the financial support awarded through the National Scholarship for PNPC and SNI programs designed for promoting quality research and close the existing gap between industry and academia.

References

1. Chintan Amrit and Joanne ter Maat. Understanding Information Centrality Metric: A Simulation Approach. (December 2018), 2018.
2. K Dharmarajan, Farhanah Abuthaheer, and K Abirami. Sentiment analysis on social media. 6:210–217, 03 2019.
3. Li Huan, Peng Richard, Shan Liren, Yi Yuhao, and Zhang Zhongzhi. Current flow group closeness centrality for complex networks?. pages 961 – 971, 2019.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32

4. 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.
5. Yoon Kim. Convolutional Neural Networks for Sentence Classification. *CoRR*, abs/1408.5, 2014.
6. Young Seok Kim and Mijung Kim. 'A wisdom of crowds': Social media mining for soccer match analysis. *IEEE Access*, 7:52634–52639, 2019.
7. Raghvendra Kumar and Prasant Kumar Pattnaik. *Graph Theory.* Laxmi Publications Pvt Ltd, 2018.
8. Giuseppe Liotta, Roberto Tamassia, and Ioannis G. Tollis. *Graph algorithms and applications 4.* World Scientific, 2006.
9. Adela Ljajić, Ertan Ljajić, Petar Spalević, Branko Arsić, and Darko Vučković. Sentiment analysis of textual comments in field of sport Sentiment analysis of textual comments in field of sport. (November), 2015.
10. Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.
11. Fabián Riquelme, Pablo Gonzalez-Cantergiani, Xavier Molinero, and Maria Serna. Centrality measure in social networks based on linear threshold model. *Knowledge-Based Systems*, 140(January):92–102, 2018.
12. Yannick Rochat. Closeness centrality extended to unconnected graphs: the harmonic centrality index. 2009.
13. 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.
14. 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.
15. Lee Wei-Meng. *Python Machine Learning.* Wiley, 2019.
16. Grace Yan, Nicholas M. Watanabe, Stephen L. Shapiro, Michael L. 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*, 19(4):419–436, 2019.
17. Junlong Zhang and Yu Luo. Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network. 132(Msam):300–303, 2017.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77980-1_32