Do popular opinion and temporal statistics interfere in FIFA soccer players ratings ?

Lucas Félix
Univesidade Federal de Minas Gerais
Brasil
lucas.felix@dcc.ufmg.br

Marcos Gonçalves
Univesidade Federal de Minas Gerais
Brasil
mgoncalv@dcc.ufmg.br

Jussara Almeida
Univesidade Federal de Minas Gerais
Brasil
jussara@dcc.ufmg.br

ABSTRACT

Soccer is currently the most popular sport in the world, and due it popularity it attracts millions of peoples to watch and practice the game. Nevertheless, soccer does not have only the attention of common people, it also attracts the recognition of researchers in many fields from sociology to computer science. Consider the field of computer science, soccer studies spreads to many areas from optimization to machine learning. Literature works related to soccer until 2013 mostly used statistics to make their evaluation. However, beginning at 2014 works start using FIFA video game data to make their evaluations. FIFA is a soccer simulator famous in whole world for it similarity with the reality, be in the players appearance or the skills model. Data from FIFA is used for been summarized, less sparse and more easy to treat than players several statistics per game. However, it has the disadvantage of been a black box. This way, in this work we tackle the problem of verifying if FIFA data is trustworthy and accurate with the reality. As ground truth we use players statistics per match and make a correlation between them and FIFA related features. Also, we verify if the public opinion reflects in a players skills set of FIFA, by collecting comments about the players evaluated and using sentiment analysis over it. Our results shows that, few players and pair of features have correlation (positive or negative). Also, we verify that, the public opinion do not interferes in a players overall. However, it can be used as possible indicator of what will be a player overall.

KEYWORDS

Sports Analytics, Data Mining, Machine Learning, Soccer

1 INTRODUCTION

Recently, the sports industry has surrender to data driven analysis in the field and business side, making data analysis a vital part of teams decision-making steps. Thus, with the premise that more information can assist win championships, teams, clubs, coach's, managers and athletes search ways to evaluate and improve their performance, given that more winnings implies in increase in the number of fans and monetary revenue [18].

Data driven solutions are applied in sports from the prediction of matches in Soccer and Martial Mixed Arts (MMA) [10, 35], passing to more analytic evaluations [38]. In resume, data is everywhere in the sport field [18].

Considering a global context, soccer is the most popular sport in the world [26]. Different from other regional sports as Baseball in United States and Rugby in Australia, soccer attracts fans in every part of the globe. Due this fact it has many practitioners and attracts enormous amounts of people to events as FIFA World Cup, continental and intercontinental championships, been often between most discussed topic in social networks as Twitter [2], having one of the most liked photos on Instagram [3], and most liked pages on Facebook [4].

In financial terms, soccer has one of the biggest monetary movement between all sports [7, 8], produced by ticket sells, TV contracts, marketing and merchandising. Illustrating that, only in Europe in 2016/2017 season, soccer has moved approximately 25 billions of euros [14]. Also, there is a financial compensation attached to clubs, like uniform sells, sponsors, TV channels quotas, also the revenue of players transfers [41].

For it popularity, soccer attracts not only the attention of fans and common people, but it also attracts the attention of researchers in several fields. Considering the range of scientific papers studying soccer it is possible to notice the presence of soccer in sociology [24], physical education [40], mathematics [32], economy [27], statistics [12] and computer science [13].

When considering the computer science branch soccer studies also spreads to many study areas as Optimization [28], Data Mining [13], Machine Learning [34] and Complex Networks [41], with studies ranging from soccer transfers optimization to predicting winnings in soccer matches.

Former soccer studies in the computational field mostly use soccer matches and players statistics to made their evaluations. Nevertheless, more recently works have been widely applying FIFA video game data in their evaluations. FIFA electronic game is a soccer simulator produced by Electronic Arts Games (EA), with major success around the globe.

FIFA is known for it fans by the realism when modeling soccer players appearances and skills. To measure the skills and evaluate a player, EA sports use games statistics and several reviewers are responsible of watch players and plays, and give scores to athletes [1]. Moreover this data counts with 800 teams and 18000 players across many countries [5]. It data has as advantage the fact that it summarizes a player features, is periodically released and it is more easy to treat than statistics per match (which is more sparse). However, some disadvantages are present in FIFA games data, given that does not comprise all the players and professional leagues in the world, having only accredited leagues, and there is also the fact that the scores of a player is mostly linked on the reviewers perception. However, it is not know how trustworthy and accurate this data are, been the modeling a black box.

Given this facts this works proposes a empirical study over the FIFA data. Our aim is to verify if FIFA data is trustworthy and accurate by comparing it with players statistics per match. With this analyze is possible to assist/guide new researchers when making soccer evaluations. More specifically, in this work we tackle the problem of (i) evaluating the trustworthy of FIFA game data, and analyze if (ii) public opinion also implies over the player skills. In the end, we want to know if FIFA data can be used in scientific researches or it is better to use statistics.

To accomplish this work we gather data of 50 top soccer players. Our criteria to select the athletes was the number of followers on Twitter [6], given that more player popularity means more data. To evaluate FIFAs data trustworthy we first assemble a dataset compose by several data sources as (i) FBref [1], which has players temporal statistics available from 2015 – 2019, (ii) SOFIFA [2], which has FIFA data from 2007 – 2019 including update releases, and also has users comments about the players.

In the end our research questions are:

(1) FIFA features trustworthy and accurate to the reality (statistics) ?
(2) Public opinion is also taken in consideration when measuring the FIFA features ?

To answer the research questions, we propose a methodology based on temporal correlation between statistics and FIFA features, and to evaluate the public opinion sentiment analysis was used. Our results shows that there is a percentage of related features between FIFA and the statistics, however, is not a representative percentage, been most of the features unrelated. When considering the public opinion a correlation between it and the general FIFA features is also nonexistent, showing that the public opinion is do not taken in consideration when calculating the features in FIFA. Our hypothesis, which will be verified in feature works, is that FIFA features are a linear combination of many statistics.

2 PROBLEM STATEMENT

Our target problem in this work is to evaluate the trustworthy of FIFA game series data by comparing it with statistics from players and public opinion collected from social media.

Thus, be $X_pe_r = X_p1e_r + X_p2e_r + ... + X_pNe1_r$, be the set of FIFA features N of a player p during in a game edition e, and a release r. And $S_pt = S_p1t + S_p2t + ... + S_pKt$ is the K statistics of a player during a period t.

Hence, we want to evaluate if there is a relation between $S_pt \rightarrow X_pe_r$, studying the trustworthy of FIFA data set X_pe_r.

3 RELATED WORKS

Given that our aim in this work is to verify if FIFA game data is trustworthy, and if public opinion is befitting with in FIFA data, in this section we verify works related to ours in two way: (i) Works that use FIFA Data 3.1, and methodologically using (ii) Sentiment Analysis 3.2.

3.1 FIFA Data

FIFA is a game series produced by EA Sports, been a soccer simulator released each year since 1993. Since then FIFA has sold over 260 million copies, and attracts each day more and more players. However, studies using FIFA game data has began more recently, in 2014 by [34], where the author aim to predict soccer matches using machine learning techniques. From that, works as [31], have also proposed a framework to predict soccer matches using FIFA data and machine learning.

Since then, more and more works have analyzed soccer using FIFA data. The second work of [13], uses FIFA data to make a qualitative analyzes of different game styles. In this work is evaluated teams that has different game styles, prioritizing different skills. In [21], the authors characterize FIFA data attributes and proposes a methodology to identify the evolution in the skills of soccer players this data source.

In [39], the authors present a approach for predicting the potential of professional soccer players. The study of [42] uses the estimated salary given by FIFA data, joined with the players skills to train several supervised machine learning techniques to estimate the wage of a soccer player.

In [25], statistical techniques are applied over FIFA data. The authors use regression method to evaluate the contribution of each player to a team victory. The approach is called Adjusted Plus-Minus (APM), been applied to other sports as hockey and basketball. The work of [30] also estimates the contribution of a player in a match. The main difference from [25] is that a framework is proposed, using data of 20 thousand European championships matches, and FIFA data, assisting in the forecast of a team win probability.

The works of [28, 29], proposes optimization models to maximize the efficiency and efficacy of a team. The work of [28], proposes a stochastic model that aim to ensure the required skills set needed to compose a strong team, respecting constrains of regulations and budget limit. The following work of [29], proposes a more complete linear programming model, using FIFA data players features. The author uses FIFA data to estimate the value and the salary of a player using Simple Moving Average method.

Lastly, we present the work of [42], which does not use FIFA data, however, aim to verify how the opinion of journalists and public attention are correlated with the price of a player and evaluate this value. The build dataset has a set of 10 features considering internal and external, whichever does not have a considerable number of instances to predicted the athletes price, as pointed by the authors. Nevertheless, this paper does not consider a set of important variables that can interfere in the players price, and do not make clear how they evaluate the reviews made by fans and journalists.

3.2 Sentiment Analysis

Sentiment Analysis (SA) consists on the task automatically detect polarity (positive, neutral, negative) of a text corpus. This technique has been applied in many fields to model, understand the behavior of users and evaluate opinions in texts [33]. Currently in the literature two different approaches has been used to sentimental analyzes: Machine Learning based (ML) and Lexicons based.

The main advantage of ML based approaches is it ability to fit within a specific context, given that a model is trained for particular purpose. However, this kind of technique flaws for it need of labeled data, which is highly costly. Recent efforts in ML based SA, use deep learning methods to infer the sentiment polarity of text corpus [15]. Although it usually achieve better results them classic ML algorithms, deep learning techniques mostly has the same issue as ML proposals. Solutions which apply ML approaches are

[1] fbref.com
[2] sofifa.com

sensible to the context of a given entry, i.e., for each training base a re-validation of the classifier is necessary, where a new a dictionary of labeled data must created. Given ML based disadvantages, recent literature efforts proposes Lexicon based methods.

Lexicon based methods create and combine different natural language processing techniques to determine a sentence polarity [23]. There are three different strategies used to create lexicons: (*i*) dictionary based, (*ii*) corpus-based and (*iii*) manual [22].

(*i*) Dictionary Based approaches use a small word seed set in online dictionaries as WordNet and Thesaurus, searching for synonymous and antonyms of the words [16]. Synonymous of positive words from the seed set are positively defined, antonyms are defined as negative. The same process happen for the negatives words from the set. Then, the found new words feed the search for new synonymous and antonyms, and the process is continuous while new words are found [22]. (*ii*) Corpus-Based methods rely on the main idea of find opinion words on large text corpus. Firstly, a seed of adjectives list is defined, then this seed set is used to find new opinion words. However, this kind of approach cleverly, use a set of linguistic constraints and conventions on connectives to find new words and it orientation [20]. Lastly, (*iii*) Manual approaches, consists in expanding human curated-lexicons (dictionaries). VADER [19], is currently the state-of-art in sentiment analysis using lexicons. This method bases in several grammatical and syntactical rules grammatical of English language, based in different literature proposals as SentiWordNet [16], Linguistic Inquiry and Word Count (LIWC) [37], Affective Forms of English Words (ANEW) [11] and General Inquirer [36]. In addiction, the authors also made a human curated dictionary with words polarity rating.

One drawback of lexicon based approaches is the need of different lexicons for each language. In some literature works, a translate step is done [9], however some semantic characteristics are lost in automatic transcribe.

Lastly, we high value that none of this works have study the validation and trustworthy of the FIFA data. Thus, our work, to the best of our knowledge is the first work to explore this venue, and can serve as ground for future works.

4 PROPOSED STRATEGY

In this section, we present our proposed methodology, given that our aim with this work is to (*i*) evaluate the trustworthy and accuracy of FIFA Game Data, analyzing how much it represents of players statistics, and verify (*ii*) if the public opinion is relevant to define the statistics of a soccer player. We propose a methodology that can be decomposed in three steps major steps: (*i*) Data Gather, (*ii*) Temporal Correlation, (*iii*) Sentiment Analysis. How several methods could be applied in steps (*ii*) and (*iii*), we only describe the approaches used in this work, justifying the use of each method. Lastly, we high-value that other methods are used in our analyses, and it are presented in the section of Experimental Evaluation 5.

4.1 Data Gather

To accomplish this work, the amount of evaluated athletes was narrowed to the 50 top soccer players on Twitter by followers. Limiting the amount of players it was possible to make a more detailed and qualitative analysis. From this, it was collected two different sources of soccer data SoFIFA and FBRef. Bellow we introduce the data present in each one of the websites and how it were used in our proposed methodology:

- **SoFIFA:** It is a major database from the game FIFA. The page has data from 2007 until 2020 FIFA editions, having also updates and users comments about athletes. The website holds the information periodically, having for each player all his features updates. Besides that, the website has for each player a module for users interaction where fans can make comments about a specific player. In this work, we apply the FIFA collect dataset comparing with a ground truth data (Statistics Dataset - from FBRef). This way it is possible to verify if the FIFA data is accurate to be used in literature works. Besides that, was also collected the comments about each player aiming to verify if the public opinion is also taken in consideration when measuring FIFAs attributes data [3].
- **FBRef:** It is a database for soccer players statistics. The page has for each player matches statistics from 2015 − 2019. As pointed before, this dataset will be used as ground truth, as it reflects the reality of the player, in comparison with FIFA data [4].

4.2 Temporal Correlation

A correlation is a possible relation between random variables. In a temporal context, the correlation between two features measures the evolution of assumed values along time [17]. In this work, we use Spearman Rank Correlation Coefficient to verify how FIFA features and soccer players statistics are related. The choose of Spearman's correlation is justified by the fact that different from Pearson's correlations that assumes that the variables are linear dependents, the metric used in our work does take this in consideration for measuring a rank difference, hence, been best suited for our problem.

The correlation made in our work is performed by window. A window w, is defined as the period between a FIFA release $X_p e_r$ and $X_p e_{r+1}$. Within this window, all statistics are aggregated $w = \sum_{t=r}^{r+1}(S_p t)$, where the time slice t most be $r + 1 > t \geq r$. It is due to notice that t and r, represents timestamps from when a player soccer match happen and when a FIFA update was released, respectively, in section 2 all this values are defined. After defined w, the Spearman correlation is measure between $\rho(w, X_p e_{r+1})$. In resume, we are aggregating all statistics matches between FIFA releases and them correlating them with a posterior release. Through Equation 1, it is possible to see how Spearman correlation is measure between the features.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

Where, n is the amount of observations, $d_i = rg(X_i) - rg(Y_i)$ is the post difference between two observations X_i and Y_i.

4.3 Sentiment Analysis

In order to evaluate the sentiment over textual features in the collected dataset, was applied the *Ifeel* strategy [9]. The *Ifeel* proposal is a

[3] https://www.github.com/lucasgsfelix/SoFIFA-Crawler

[4] https://github.com/lucasgsfelix/FBRef-Crawler

benchmark of sentiment analysis with 19 different *state-of-practice* sentiment analysis techniques. The general aim of this strategy is to facilitate the task of polarity evaluation over text corpus.

Most important, *Ifeel* provides also a multilingual support for over 60 languages, assisting in analyzes made in social networks where the comments can be from anywhere in the globe. Thus, the algorithm works by performing the translation of any language to English, and then performing Sentiment Analysis applying techniques as *VADER, SentiWordNet, SentiStrength* and others over the documents.

We high-value that all algorithms present in *Ifeel* perform in a sentence level, without requiring prior knowledge of users opinion to make inferences about the collection of documents. The disadvantage of this proposal is it high cost due the several implemented methods. However, this is offset by the possibility of choose the implemented algorithm within *Ifeel* that better fits the dataset.

When making this analysis it was possible to notice that VADER was the method that better separate the sentiment over the comments. Thus, in our final analysis VADER was used.

Lastly, we high value that we make available all datasets and code implemented to perform this work.

5 EXPERIMENTAL EVALUATION

In this section, we evaluate how the databases collect to make accomplish this work are related. First, we describe the datasets. Next, we report and analyze the results of correlation obtained between the datasets.

5.1 Datasets

To accomplish this work we collect several different datasets. Each one of them holds different information about the soccer players, as pointed in Section 4.

From SoFIFA was collect the features set that compose a player and all it updates. This features summarizes the abilities of a soccer player been mostly represented by numerical values, measured by the producer of the game FIFA. The assembled dataset have for each player his features set and periodically releases. FIFA data is normalized between 0 − 100 and it is a complete matrix, thus, even player that does not act in certain position holds characteristics specifics from other position. One example is, even tough Cristiano Ronaldo does not play as Goalkeeper he have the abilities of a goalkeeper. Worth mentioning his abilities as goalkeeper are much lower than a player that acts in the goal. The same way, goalkeeper has the same set of features that a midfielder player, however not with the same quality.

From FBRef was collected statistics of a player per match. This features set have all plays that a athlete have made per game as number of goals, passes, crosses, cards taken. It is high-valued that all features are numerical.

In SoFIFA comments, was collected for each player the comments leaved by users in SoFIFA platform. This comments are comparing, good or bad talking about a player skills in real life and in the game. In the comments there are textual and non-textual documents. On the developed web crawler only textual features were considered, given their facility, when compared to other multimedia features, to treat and evaluate.

Dataset	# Rows	# Columns	Mean	Period
SoFIFA	1299	69	28 Updates	09/2014-08/2019
FBRef	8939	33	194 Matches	02/2014-06/2019
SoFIFA Comments	198168	6	4608 Comments	12/2012-10/2019

Tabela 1: Datasets Information

In Table 1, a brief summary of the used datasets: number of rows and columns, the mean by the amount of players and period present in the dataset.

We high-value that to proper correlate the features, we manually separate the sets statistics and FIFA features to be compared. This way, we avoid to vainly measure correlation of features that is known to not have relation. Figure 1 shows features correlated of FIFA (in green) and FBRef (in yellow).

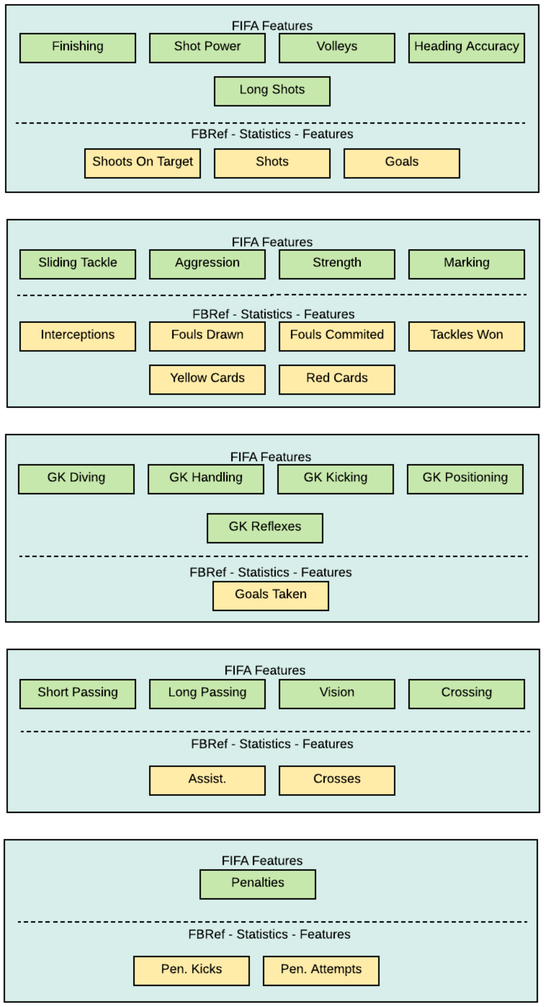

Figura 1: Sets of statistics and FIFA features separation

5.2 Statistics x FIFA Features

To answer our first research question *"FIFA features trustworthy and accurate to the reality ?"*, we perform a temporal correlation analysis between FIFA Features and statistics collected from FBRef. The Figure 2, shows the heat map with the correlation for all players with at least one valid correlation. When looking at map it is possible to see that few blocks shows correlation (positive or negative) (dark blue or close to white). To better evaluate the data and verify if there is correlation between the features, we verify the distribution of each pair of features which is possible to see in the Appendix Section 7.

When evaluating the pair-wise feature distribution, two of the graphs shows more attention for the possibility of distinguish players with a high correlation, been the Figures 6, 9. Through the Figures it is possible to notice the distribution for the 10° 6 and 90° 9 percentile. The graphs are composed by the 10° percentile and 90° percentile for each pair of features between the players correlation set. Considering that correlation values smaller between $-0.3 < \rho < .3$, do not show correlation negative or positive, we select pair features and players that appears outside this range that are in the 10° and 90 percentile. From this values we evaluate players and pair of features that presents the most positive and negative correlation.

When evaluating the players within the 10° percentile it is noticed that there are players with more presence between the players with smaller pair-wise correlation. Between those there are Pedro, Radamel Falcão, Bastian Schweinsteiger, Gareth Bale and Zlatan Ibrahimovic, been 24 of the total amount of players inside this percentile. In Table 2, it is showed the most frequent players and pair-wise features within 10-th percentile.

Player	Appear.	Min(ρ)	Max(ρ)	Min Feat. Pair.
Pedro	6	−0.729	−0.605	Mark. - Tac. Won
Radamel Falcão	4	−0.836	−0.40	Aggres. - Red C.
Bastian Schweinsteiger	4	−0.836	−0.751	Long Shots - Shots
Gareth Bale	4	−0.646	−0.479	Mark. - Fouls Commit.
Zlatan Ibrahimovic	3	−0.6	−0.448	Crossing - Assist.

Tabela 2: 10-th percentile description

Table 2, shows for each of the top players with most appearances within 10° percentile his min and max correlation, and the pair of features corresponding of the lowest correlation. The values within this quantile have mostly negative correlation, showing that 90% of the correlation values of that pair feature are bigger than that pointed.

When evaluating the players and the pair of features, it is noticed that the lowest correlation for three of the five players are related with characteristic that not are from his original position, given that Pedro, Radamel Falcão and Bastian Shweisteiger acts from the middle field forward, but the pair of features represents defense characteristics as Marking - Tackle Won, Aggression - Red Cards, Marking - Fouls Committed. This can show that FIFA data tend to represents better characteristics of the position of the player, thus, this players are poorly represented for acting in another field area having a negative correlation, even tough their statistics showing that they are good in defense areas.

Player	Appear.	Min(ρ)	Max(ρ)	Max Feat. Pair.
Pedro	5	0.461	0.717	Vision - Crosses
Arturo Vidal	4	0.461	0.791	Short Pass. - Crosses
Mesut Özil	3	0.443	0.691	Strength - Fouls Drawn
Luis Suárez	3	0.734	0.855	Agress. - Fouls Drawn
Radamel Falcao	3	0.643	0.681	Volleys - Shots

Tabela 3: 90-th percentile description

When evaluating the 90° percentile it is possible to notice that the most correlated players act from midfield forward. Within this percentile there are 28 players, and the values shows players and pair of features most correlated. It is possible to notice that the pair of features most correlated are mostly characteristics from strikers and midfielders, represented by Pedro, Arturo Vidal and Radamel Falcão. Luis Suárez has a high aggression values for his current bad behavior in games [5].

Most of the players shows correlation only when considering a small amount of features. Nevertheless, some players as Pedro and Radamel Falcão are present both in the list of players with positive and negative correlation, which evidences that features are better modeled given the position of the athlete. Hence, even tough a players help the defense, his defense attributes will not increase, given that this are features do not represent the characteristics of his original position.

5.3 Popular Opinion Evaluation

To evaluate the popular opinion, we focused on the correlation between the sentiment of SoFIFA comments and the overall of a player. As the FIFA Features, was made a temporal correlation of the overall and the sentiment. The overall of a player is the weighted mean of all his features. The weights give more importance to features directly related to the players position. Thus, a goalkeeper have as mainly features GK Kicking, GK Reflexes.

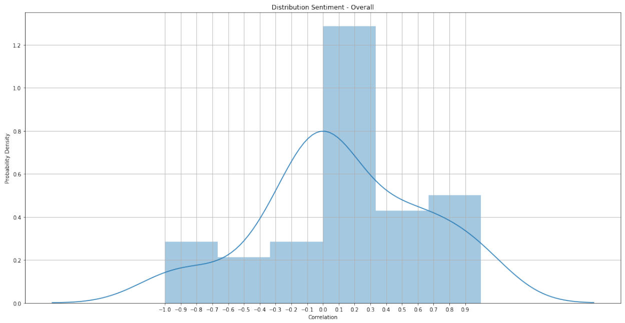

Figura 3: Sentiment x Overall Correlation Distribution

The Figure 3, shows the correlation between sentiment and overall distribution. Of the total amount of points, 42, 21 are in correlated zones were $\rho \leq -0.3$ or $\rho \geq 0.3$. The points represent the correlation between the sentiment and overall in a window w. Thus, for each player there is a correlation between his comments sentiment and each release overall, in total there are 232 points but

[5] https://www.theguardian.com/football/blog/2018/mar/27/world-cup-stunning-moments-luis-suarez-bites-giorgio-chiellini-in-2014

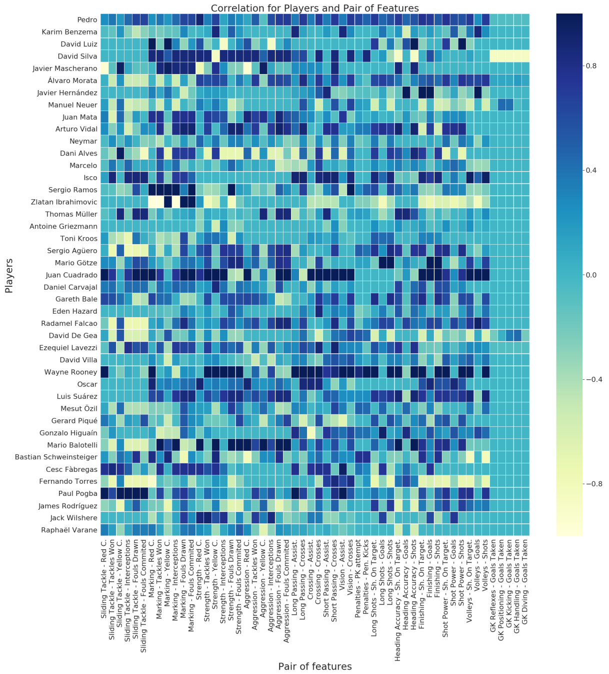

Figura 2: Correlation between players and pair of features

only 42 points of correlation were considered, thus less than one per players (50 in total). In other words this represents that for all players and releases, comments in the release window does not show any correlation (not even a number) with the overall, been sign of the lack of correlation between the popular opinion and the players abilities.

In conclusion, answering our second research question *Public opinion is also taken in consideration when measuring the FIFA features ?* the answer is no. Even tough some players show a positive correlation, most of them have no correlation at all. Players positive correlation with the overall can be just a sign that the athlete has playing good in real life, thus, as the positive comments increase, also will increase the overall of the player, the same holds for negative comments and decrease of the overall.

6 CONCLUSION

In this work, we evaluate FIFA features dataset, which has been widely applied in literature works to model and make predictions over soccer studies. Our aim with this validation is to verify if the FIFA features are trustworthy and accurate. To accomplish this work we use statistics data, FIFA releases data and users comments about the players. In our methodology we use Spearman as temporal correlation and iFeel to sentiment analysis, aiming to answer our research questions.

Our results show that most of the features and players does not have correlation with the statistics. With this results we answer the first question about the accuracy and trustworthy of FIFA data: FIFA data does not follow the statistics, been poorly accurate of the reality of the players. Our results show that, independent of how the statistics increase or decrease, generally this does not interfere in FIFAs abilities. Nevertheless, was possible to notice that FIFA miss more in abilities that are not directly related to original position of the players. Thus, if a striker assists the team marking players, this will not taken in consideration, even tough his statistics shows that he is good at it.

Answering of second research question, if the popular opinion makes any difference in general features of the player, our results shows that also there is no correlation, thus public opinion is not considered, independent if is good or bad talk about the player. The public opinion can only reflect what will be overall, will increase or decrease, however, it does not directly interfere in the total punctuation. Also, it was possible to analyze that FIFA releases tend to varies little to accurate simulate real world players.

As future works we intend to collect more data: with features statistics dataset, more comments and from another sources, as Twitter to represent the public opinion. To better evaluate the public opinion we aim to performance temporal topic modeling over the comments summarizing the comments and evaluating the sentiments over the topics. Lastly, we aim to propose a metric that can better adapt to players statistics and give the players a more real features, assisting future literature works that intend to evaluate soccer.

REFERÊNCIAS

[1] [n. d.]. EUROGAMER, How EA determines FIFA player ratings. https://www.eurogamer.net/articles/2016-09-27-how-ea-determines-fifa-17-player-ratings. Accessed: 2017-11-17.

[2] [n. d.]. Here are the most retweeted sports tweets of 2016. https://www.si.com/extra-mustard/2016/12/06/sports-twitter-most-retweeted-tweets-moments-2016. Accessed: 2019-08-20.

[3] [n. d.]. Here are the most retweeted sports tweets of 2016. https://en.wikipedia.org/wiki/List_of_most-liked_Instagram_posts. Accessed: 2019-08-27.

[4] [n. d.]. List of most-followed Facebook pages. https://en.wikipedia.org/wiki/List_of_most-followed_Facebook_pages. Accessed: 2019-08-27.

[5] [n. d.]. Meet the Data Master behind EA Sports' Popular FIFA Franchise. https://datamakespossible.westerndigital.com/meet-data-master-ea-sports-fifa/. Accessed: 2019-08-29.

[6] [n. d.]. Most followed soccer players on Twitter. http://www.tweetsfc.com/stats/most-followed-football-players-on-twitter. Accessed: 2019-09-10.

[7] [n. d.]. Soccer falls short from being the sport with the highest revenue. https://www.thenewbarcelonapost.com/en/soccer-falls-short-from-being-the-sport-with-the-highest-revenue/. Accessed: 2019-08-29.

[8] [n. d.]. Soccer global dominance in three simple charts. https://www.mic.com/articles/91009/soccer-s-global-dominance-of-sports-in-3-simple-charts. Accessed: 2019-08-29.

[9] Matheus Araújo, Pollyanna Gonçalves, Meeyoung Cha, and Fabrício Benevenuto. 2014. iFeel: a system that compares and combines sentiment analysis methods. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 75–78.

[10] Rahul Baboota and Harleen Kaur. 2018. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting* (2018).

[11] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Citeseer.

[12] Julen Castellano, David Casamichana, and Carlos Lago. 2012. The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of human kinetics* 31 (2012), 137–147.

[13] Leonardo Cotta, POV de Melo, Fabrício Benevenuto, and Antonio AF Loureiro. 2016. Using fifa soccer video game data for soccer analytics. In *Workshop on Large Scale Sports Analytics*. XX.

[14] Deloitte. June 2016. Annual review of football finance.

[15] Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 69–78.

[16] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining.. In *LREC*, Vol. 6. Citeseer, 417–422.

[17] António Jorge Teixeira Falcão. 2012. *Detecção de Correlação e Causalidade em séries temporais não categóricas*. Ph.D. Dissertation. Faculdade de Ciências e Tecnologia.

[18] Gil Fried and Ceyda Mumcu. 2016. *Sport analytics: A data-driven approach to sport business and management*. Taylor & Francis.

[19] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

[20] Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*. Association for Computational Linguistics, 174–181.

[21] Ivan R. Soares Jr., Renato M. Assunção, and Pedro O. S. Vaz de Melo. 2018. Entendendo a evolução das habilidades de jogadores de futebol através das pontuações do jogo eletrônico FIFA. *The Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)* KDMILE (2018).

[22] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 415–463.

[23] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1909–1918. https://doi.org/10.1145/3178876.3186168

[24] Joe A Maguire. 1988. Race and position assignment in English soccer: A preliminary analysis of ethnicity and sport in Britain. *Sociology of Sport Journal* 5, 3 (1988), 257–269.

[25] Francesca Matano, Lee F Richardson, Taylor Pospisil, Collin Eubanks, and Jining Qin. 2018. Augmenting Adjusted Plus-Minus in Soccer with FIFA Ratings. *arXiv preprint arXiv:1810.08032* (2018).

[26] Palacios-Huerta and Ignacio. 2004. Structural changes during a century of the world's most popular sport. *Statistical Methods and Applications* 13, 2 (01 Sep 2004), 241–258. https://doi.org/10.1007/s10260-004-0093-3

[27] Ignacio Palacios-Huerta. 2016. *Beautiful game theory: How soccer can help economics*. Princeton University Press.

[28] Giovanni Pantuso. 2017. The football team composition problem: a stochastic programming approach. *Journal of Quantitative Analysis in Sports* 13, 3 (2017), 113–129.

[29] Vineet M Payyappalli and Jun Zhuang. 2019. A data-driven integer programming model for soccer clubs' decision making on player transfers. *Environment Systems and Decisions* (2019), 1–16.

[30] Konstantinos Pelechrinis and Wayne Winston. 2018. Positional Value in Soccer: Expected League Points Added above Replacement. *arXiv preprint arXiv:1807.07536* (2018).

[31] Darwin Prasetio et al. 2016. Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. IEEE, 1–5.

[32] Hindriyanto Dwi Purnomo and Hui-Ming Wee. 2015. Soccer game optimization with substitute players. *J. Comput. Appl. Math.* 283 (2015), 79–90.

[33] Leonardo Rocha, Fernando Mourão, Thiago Silveira, Rodrigo Chaves, Giovanni Sa, Felipe Teixeira, Ramon Vieira, and Renato Ferreira. 2015. SACI: Sentiment analysis by collective inspection on social media content. *Web Semantics: Science, Services and Agents on the World Wide Web* 34 (2015), 27–39.

[34] Jongho Shin and Robert Gasparyan. 2014. A novel way to soccer match prediction. *Stanford University: Department of Computer Science* (2014).

[35] Leandro AA Silva, Johnnatan Messias, Mirella M Moro, Pedro Olmo Vaz de Melo, and Fabricio Benevenuto. 2015. Algoritmos de Aprendizado de Máquina para Prediç ao de Resultados das Lutas de MMA. *Brazilian Symposium on Databases* (2015).

[36] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis. (1966).

[37] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[38] Pedro OS Vaz de Melo, Virgilio AF Almeida, Antonio AF Loureiro, and Christos Faloutsos. 2012. Forecasting in the NBA and other team sports: Network effects in action. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 3 (2012), 13.

[39] Ruben Vroonen, Tom Decroos, Jan Van Haaren, and Jesse Davis. 2017. Predicting the potential of professional soccer players. In *Machine learning and data mining for sports analytics ECML/PKDD 2017 workshop*, Vol. 1971. 1–10.

[40] Phillip Ward, Yaohui He, Xiaozan Wang, and Weidong Li. 2018. Chinese secondary physical education teachers' depth of specialized content knowledge in soccer. *Journal of teaching in physical education* 37, 1 (2018), 101–112.

[41] Liu XF, Liu Y-L, Lu X-H, Wang Q-X, and Wang T-X. 2016. The Anatomy of the Global Football Player Transfer Network: Club Functionalities versus Network Properties. *PLoS ONE* 11, 6 (05 Feb 2016). https://doi.org/10.1371/journal.pone.0156504

[42] L Yaldo and Lior Shamir. 2017. Computational estimation of football player wages. *International Journal of Computer Science in Sport* 16, 1 (2017), 18–38.

7 APPENDIX

7.1 Statistics x FIFA Features

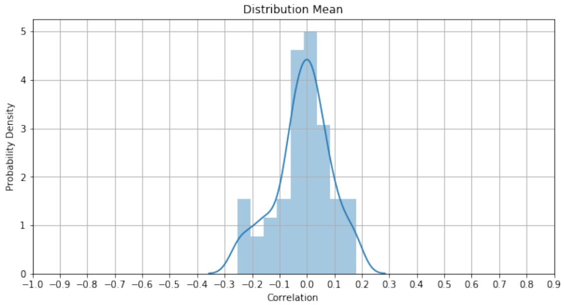

Figura 4: Players Mean Correlation Distribution

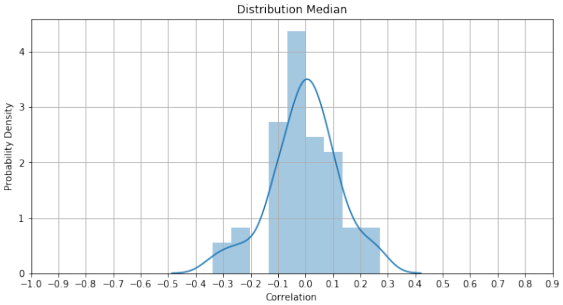

Figura 5: Players Median Correlation Distribution

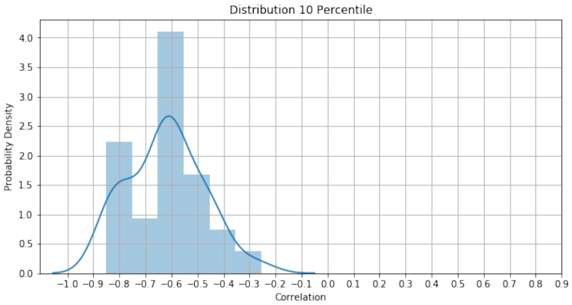

Figura 6: Players 10º Percentile Correlation Distribution

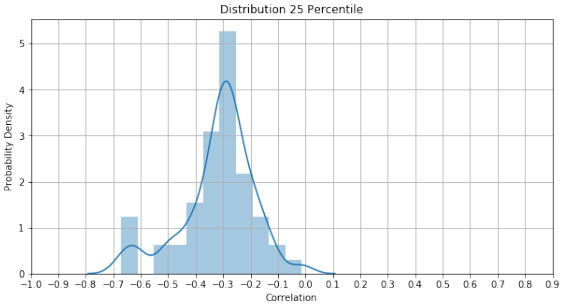

Figura 7: Players 25º Percentile Correlation Distribution

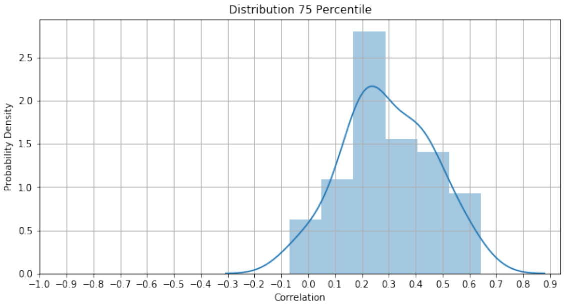

Figura 8: Players 75º Percentile Correlation Distribution

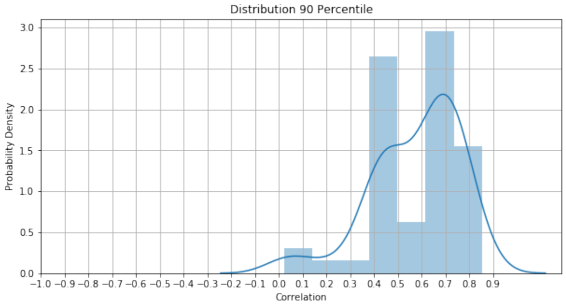

Figura 9: Players 90º Percentile Correlation Distribution