What Influences Football Enthusiasts When Setting a Player`s Market Value

Ricardo Meneses Flores

ANR: 345432

Master's Thesis

Communication and Information Sciences

Specialization Data Journalism

Faculty of Humanities

Tilburg University, Tilburg

Supervisor: Dr. E. A. Keuleers

Second Reader: Dr. R. M. F. Koolen

July 2017

Abstract

This research examines the relationship that exists between performance and external variables with the player's market value set by the crowdsourcing platform Transfermarkt.com. Sports media journalists, fans, sports agents, and managers have been using this rating as an official information source in the last decade. The present research contributes to current models by including external variables, which represent the role of the media and social media through Twitter. Also, this study compares the performance of three models; first, a performance (only talent based variables) model; second, a non-broadcast (including non-weighted variables) model; and third, a broadcast (including weighted variables) model. The analysis takes the English Premier League (EPL) 2015/2016 season as a sample. Tweets from twenty teams and seven sports media organizations are used to create the external variables. The results illustrate that a regression model with weighted variables is a better predictor of a player's market value than a model with non-weighted variables. The public recognition from teams of opposing players proves to be a significant variable that influences market value. This contrasts with the influence of the media that is not a significant factor.

Keywords

Data journalism, Football, Social media, Media, Text mining

Acknowledgments

This study is the final challenge in a year that changed my life completely. First, I want to thank my parents and my sister for their support. I was able to finish this research thanks to the guidance of my supervisor, Emmanuel Keuleers. I appreciate all his time and patience during the last six months. An especial thanks you to Krishna Srinivasan for his guidance with the econometric model, but mostly for his friendship. Finally, thanks to Andrés Meneses, my brother, and my best friend. Your motivation and support made me believe in myself and in the work I was doing.

All the effort I put into this pages is dedicated to all my friends and family. This goes especially to Julian, Marta, Cristina, Anna, Carlos, Macarena, and Maxime. I'm a better person because I met all of you. I cherish all the moments I spend with each one of you. All the talks, the jokes, and the adventures are the most valuable things I will take with me for the rest of my life. This is also to my closest friends in Ecuador. Andrea and Winnie, this is mainly for you. You kept close to me during my biggest crisis. I will never forget that.

Table of Contents

1. Introduction

1.1 Football transfer fees madness and lack of market value

Every summer, sports journalists leave live action reporting in search of the latest football players transfers deals. During those days, big stories come from gossips about who is going where. For instance, in 2016, Paul Pogba was the center of attention. The Frenchman story ended with him signing a contract with Manchester United for a record fee of 105 million euros. Year after year, the inflation rate of the transfer fees increases in a considerable manner. The behavior of the football market has had this particularity as a historical trend (Dobson & Gerrard, 1999). Many factors are behind the undisclosed negotiations of player's financial movements. Team executives and player agents discuss contract terms, such as, the final price, in an effort to set a value to unique abilities and characteristics of a footballer.

Clubs, agents, and players are the main characters in the transfer market. However, for the public, including journalists, the procedure that determines the value of a player is mostly unknown. It may seem that the quality of a player is obvious, yet agents affirm that it is not based on an exact formula. Phil Smith, a player representative, described the negotiations as "a question of supply and demand, the greater the competition, the higher the price" (Foster, 2016). In a similar manner, a research conducted by Garcia Del Barrio and Pujol (2007) concluded that stronger brand names help explain higher transfer fees. Consequently, fans and the media are always expectant to know how much higher the next record fee will be and who is involved.

The football labor market is within the frameworks and norms of the International Federation of Association Football (FIFA), regional and national football associations. Only registered professional clubs can buy player's services after signing a contract. Similarly to other labor markets, the Treaty of Rome determines several norms, for example freedom of contract (Antonioni & Cubbin, 2000). In football, this entails that players can move freely between clubs. Nevertheless, none of these regulations include a public or an official method that sets up a player's market value. As a consequence, several

experts and football fanatics have designed their own ratings and models to set a player approximate value. An example of this is Transfermarkt.com, an online crowdsourcing platform that collects data from football players. The information of this website will be used in this research paper and its role further explained in the Methodology section.

1.2 The coverage of football players' transfers and market value

The rise of football as the most popular sport in the world comes with an extensive media coverage. Along with the development of the game, sports journalists and commentators have built communicative products to analyze the performance of players. Because of daily coverage of football related topics players have been transformed into idols, celebrities, and stars figures. For footballers, fame comes along with ongoing and new contract deals. Players are not only considered professional athletes, but name brands. Consequently, the role of media is between informing and promoting either a positive or negative perception of a player (Haynes, 2007).

In the last decade, social media has allowed footballers to keep personalized channels to connect with the public (Hutchins, 2011). Different platforms, such as Twitter and Facebook, serve as opinion tanks before, during and after every match. Nowadays, everything that happens on a football pitch is meticulously recorded. Fans are one of the actors that contribute to the flux of information, they are neither inactive nor on hold for media to lead them, but the other way around. Users of social media are now judging and providing critiques at the pace of the media. Moreover, due to the social status of football players, news from off-field events are also part of the conversation.

1.3 Problem Statement

The connection between football players' market value and media is the brand value. The image of a player is created around its persona that can attract and relate to millions of football fans. Supporters and journalists are now reacting at the same speed thanks to social media. In contrast to what happened in the past, when sports commentators and reporters provided information and their opinions, nowadays,

footballers must deal with an even larger public coverage. From a communicational point of view, players and teams have now enough power to build their image through different platforms.

According to Stefan Dombert, an executive at transfermarkt.com, players' market values "are estimated in a rather qualitative than quantitative process" (Dombert, 2017). There are several criterions contributing to the market value. The main ones are "performance data, age, position, club, league, national team, transfer fees paid so far, possible transfer fees in future, marketing-related factors ('prestige') and future perspectives" (Dombert, 2017). In this context, media reporting is a consequence of a positive or negative development of the above-mentioned criterions contributing to the market value. Under this premise, the media impact is an effect of the market value, hence to transfer fees. Nevertheless, according to Herm et al. (2014), "community evaluations (like transfermarkt.com) can largely be explained by an econometric model that contains two blocks of determinants: variables that are directly related to players' talent and variables that result from judgments by external sources (e.g., journalists)". Therefore, the present paper intends to add valuable information to the current literature by focusing on the second group of variables mentioned by Herm et al. (2014). Similarly to the latter, we will take into consideration Brunwik's lens model (see Figure 1), a concept that establishes the premise "that observers do not rely on all possible cues when making judgments about their individuals or objects, but they rely instead on selected, probabilistic cues or attributes" (Herm et al., 2014). The lens model allows us to experiment on how media and clubs became part of the judgment in respect to a players' market value.

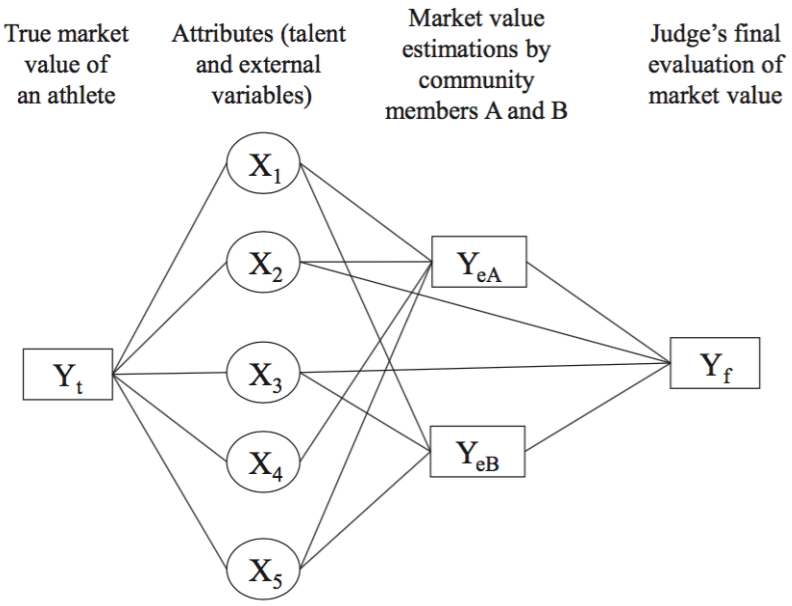

Figure 1. Brunswik's lens model of decision making for team decisions. (Herm et al. 2013)

What influences football enthusiasts to set a players' market value? We believe the media is a key factor to answer this question. Our analysis is divided into two parts. First, find if media, represented by social media measures, has a strong relationship when predicting the market value of a footballer. Second, compare the results of weight and non-weighted word frequency. We named these two conditions as a broadcast effect (weight) and non-broadcast (non-weighted) effect. Following the example of existing theoretical work (E.g. Herm et al, 2013; Herberger & Wedlich, 2016, Dobson & Gerrard, 1999), we expect to analyze how strong is the relationship between media and social media coverage (of football players and clubs), performance parameters, and the market value. The present research is based on text-mining techniques, to have access to a relative version of reality by interpreting textual documents, and a quantitative analysis, based on a mathematical model to explore the relationship between variables.

The purpose of this research is to utilize a regression model that illustrates the relationship between media and social media coverage, and the player's market value rating created by football enthusiasts. The regression models proposed in the next section emphasize in prediction and not in explanation. Prediction is about anticipating, estimating or forecasting what may happen in the future

(Shmueli, 2010). The core analysis will be about understanding correlations, which does not an indicate cause and effect.

I.e., a linear regression model is employed to understand how skills and media and social media content affect the market value of a player. We will use information of the English Premier League (EPL) from the season 2015/2016. The variables include player's data (e.g. goals, field position, age, height), and tweets from English media organizations and the respective teams. Two research questions are formulated as follows:

RQ1: To what extent do external variables and performance variables relate to the player's market value?

RQ2: Is the size of the audience measured by the number of followers, a proxy of a broadcast measure, a suitable predictor to forecast the player's market value?

2. Literature Review

2.1 Media agenda setter

In previous studies, the effect of media has been discussed through several points of view. According to McCombs and Shaw (1972), "readers learn not only about a given issue, but also how much importance to attach to that issue from the amount of information in a news story and its position". Both authors refer to this as the agenda-setting concept, understood, also, as a journalistic role. The aim of the reporter and editor is to emphasize or prioritize news before publishing. This whole procedure shows how media shapes public concern (McCombs & Shaw, 1972).

The agenda-setter role is one of the many roles a journalist could assume. Other journalistic roles perceptions are the watchdog, investigative reporter, civic educator, service, infotainment, advocate, curator, loyal facilitator, and disseminator (Fahy & Nisbet, 2014; Mellado, 2014). Therefore, the content published by media outlets depends on many factors that are not only based on facts and events. Gans (1979) and Gitlin (1977) group the theoretical perspectives into five categories:

> Content reflects social reality with little or no distortion, content is influenced by media workers' socialization and attitudes, content is influenced by media routines, content is influenced by other social institutions and forces, and content is a function of ideological positions and maintains the status quo.

The above mentioned theories have a background of examples cited by Shoemaker and Reese (1996). In summary, content is permeable to the influence of different roles.

It will be a mistake to consider journalism as a perfect reflection of reality. Authors have agreed that the media does not exactly mirror reality, but that journalists re-create a new reality after a selective procedure (Lippmann, 1946). The journalistic practice in which information is selected and filtered is called the gatekeeper function. In recent published articles, the gatekeeper theory has been replaced by 'framing' theory, which emphasizes the predictive or manipulative function of media by telling its

audience how to think about certain issues (Chong & Druckman, 2007). Furthermore, the media treats negative and positive news frames differently; in a study (de Vreese, et al., 2011), negative news frames had a significantly stronger exposure than positive news. To conclude, media coverage recreates reality but it cannot be considered as a perfect reflection of it.

2.2 Football market

In this research, the football market is the road where we will search for media influence. The path will go through the economic characteristics; social factors and theories created on this topic. Beside common aspects of any market like competition, marketing, or specialization, the football market can be distinguished by the need for illusion (Barajas et al., 2010). The virtuous circle of sports, especially with football, starts when a great number of supporters get interested, consume and identify with a player or club. It continues with media reporting and producing content. Finally, if the results are positive for most of the actors the virtuous circle starts again (Barajas et al., 2010).

The sports transfer market has been studied mostly since the mid-1950's, starting with Rottenberg (1956). Since the 1990's, many econometric models have been built to predict transfer fees, using variables such as a player's performance skills, characteristics of buying and selling clubs, and other control measures (e.g. Dobson & Gerrard, 1999; Carmichael et al., 1999; Speight & Thomas, 1997). The main differences between the models are the variables that are taken into consideration and the number of transfers they take into account. Because not all footballers get to be transferred to a new team every year, a small group of scholars has used the estimated market value of players instead of transfer fees (e.g. Herm et al., 2013). This research will follow the same approach.

Market value.

Herm et al. (2013) concluded in their study that community-based market-value estimates, like transfermarkt.com, could predict actual transfer fees. The same research stated that community evaluations can largely be explained by an econometric model that contains two blocks of determinants:

"variables that are directly related to players' talent and variables that result from judgments by external sources (e.g., journalists)". One of the issues encountered by scholars who worked with transfer values was the selection bias (Carmichael et al., 1999). Authors argue that not all players are transferred in each season, so the sample data only contains players with high profiles. That is why working with market values is beneficial. "The market quotes indicate the current expected receivable amount if a transfer would be conducted" (Herberger & Wedlich, 2016). Also, in the study conducted by Gerhards, Mutz, and Wagner (2014) the market quotes from transfermarkt.com prove to be a well proxy for real market values. These findings allow us to believe that the weight or impact of media as "external source" could add valuable knowledge to the current literature.

2.3 Superstars and media

The social phenomenon of players is similar to the role of Hollywood celebrities. The relationship between the brand and the value of a player was addressed by Garcia del Barrio and Pujol (2007). They found that the highest paid players are in such position thanks to a high brand value. This concept has been recognized as the "superstar phenomenon" (Franck & Nüesch, 2012). Under this concept, a player value is driven by two factors: talent, objective performance (Rosen, 1981), and network externalities of popularity (Adler, 1985). In the research conducted by Franck & Nüesch (2012), the performance statistics of the players encompassed talent, while popularity was operationalized by the number of citations in newspapers and weekly magazines. In addition, Herm et al. (2013) included in their model decisions of team coaches or club managers and evaluations by experts. "Hence, the superstar phenomenon can be measured by more variables, not just by the number of press citations" (Herm et al., 2013).

Acclaimed players like Cristiano Ronaldo or Lionel Messi are more than athletes, they are leaders in the current market (Arceo, 2003). "They are a reference of behavior (including consumer behavior) for many audiences, mainly young people, before whom they can more easily allow the association of image

attributes. They are a prescriber of cultural values, an element of communicative globalization" (Arceo, 2003). This happens with all footballers, but not everyone receives the same social attention because of their perceived talent and brand image. It's common that goal scorers and attacking players attract more media coverage. "There is a concentration of output among a few individuals, marked skewness in the associated distributions of income and very large rewards at the top" (Rosen, 1981). Scholars called this type of labor markets as winner-takes-all because the distribution shows a few group get disproportionately more revenue (or attention) than the majority (Frank y Cook, 1995).

As has been pointed out early, the media impact is an intangible asset with immense repercussions for the entertainment business and, specifically, for the world of football (Garcia del Barrio & Pujol, 2007). In their investigation, Garcia del Barrio and Pujol (2007) estimated popularity by computing the number of web pages that refer directly to the player and the team to which he belongs. This measure was complemented with an evaluation of the notoriety each player or team had, according to the media coverage received. "Media value must be understood as the popularity and prestige, as well as the media and social impact of individuals and clubs" (Garcia del Barrio & Pujol, 2007).

2.4 Social media and sport

The approach of the present research includes Twitter as the environment where media, players, and clubs interact. Twitter is opening new and direct forms of communication between clubs and their fans. This social network is also allowing footballers to find their voice and create powerful personal brands. Twitter's utility as a microblogging platform is exploited to offer more frequent and disposable updates, including running commentaries on games (Price et al., 2013).

How footballers and clubs manage social media accounts differs from other communication politics. The Public Relations Department of each club provides rules to their own players. According to Price et al. (2013), the challenge for sports journalists is to bring order to the chaos and suggests that "perhaps this is the time for the role of journalist as gatekeeper to evolve into that of referee. The role is to

offer informed opinion against the white noise of Twitter chatter and to question the endlessly positive version of events proffered by PR and marketing professionals" (Price et al., 2013).

2.5 Wisdom of the crowd

The present research is based on the concept known as wisdom of the crowd, which states that the collective opinion of a group of individuals is more accurate than a single expert's opinion (Surowiecki, 2005). The crowdsourcing term has been used for a long time (Galton, 1907), but the arrival of Internet provided new possibilities. According to previous studies (e.g. Charness & Sutter, 2012) that compare individual versus group decision-making, there is evidence that groups produce more rational output than individuals. Wolfers & Zitzewitz (2004) study indicated that crowds perform very well in information aggregation tasks. In the same manner, estimating a player's market value is such an information aggregation task.

2.6 Broadcast effect

Twitter offers the possibility to link users by the click of a follow button. The number of followers of a user, "directly indicates the size of the audience" (Cha et al., 2010). This measure of exposure will determine the weight factor of each media and club account. Our aim is to compare a broadcast effect (weighted) versus a non-broadcast (non-weighted) effect of information extracted from tweets. Although some studies (e.g. Cha et al. 2010, Cataldi & Aufare 2014) have shown that a user with a high number of followers should not necessarily be considered an influential agent, study results do recognize that some users have a higher influence when the topic of the discussion is well defined. Summing up, by highlighting the influential factor that each club and media have upon an audience, we could set up a clearer panorama of the role of media and PR & marketing.

In the present research, the model with non-broadcast effect variables will include the occurrence of each players' name or Twitter account. For the models with broadcast effect variables, the number of followers will multiply the number of mentions in each media or club account. For instance, if a player

was mentioned 3 times by Bleacher Report, that has 244 203 followers, then the broadcast value will be (3 * 244 203) = 732 609. Of course, each player is mentioned by several accounts so the values will be added to set a total broadcast mention value.

3. Methodology

In this research, we want to determine the influence of 24 factors on the market value of the player. The sample consists of players that were part of an English Premier League (EPL) team in the 2015/2016 season. The variables considered describe each player's performance and public influence. Performance can be measured within different attributes, like goals scored or pass success rate. The information about players' skills and market value was extracted from Transfermarkt.com and WhoScored.com. Performance data was structured and in need of little edition. On the other hand, the variables meant to address public influence, like number followers, mentions or retweets required a complex text mining process.

3.1 Text mining analysis

The basic concept of data mining is to apply certain techniques to obtain valuable information that is buried in data. The applications of text mining combined with social media have resulted in relevant information from different fields. Its main purpose has been to gather opinions about topics like climate change (Cody et al., 2015), presidential approval rates, customer sentiment (Cody et al., 2016), or referendum results (Celli, Stepanov, Poesio, & Riccardi, 2016). Text provides scientists with unstructured data that needs to be ordered and made readable for statistical software.

Figure 2 summarizes the framework used in text mining. The process involves information retrieval, text analysis, information extraction, predictive modeling, visualization, database technology, and data mining. In this research, the text that will be analyzed is presented in the form of "tweets". Tweets are shared messages, restricted to 140 characters that can include links to external websites, videos or images ('FAQs about Twitter', 2017). The intermediate form "can be semi-structured such as the conceptual graph representation, or structured such as the relational data representation" (Tan, n.d). It is the group of datasets created for each account. Finally, the knowledge distillation deduces patterns or knowledge from

the intermediate form (Tan, n.d.). In this research, the knowledge distillation is the count of player's names or Twitter accounts along with the performance values in the final dataset.

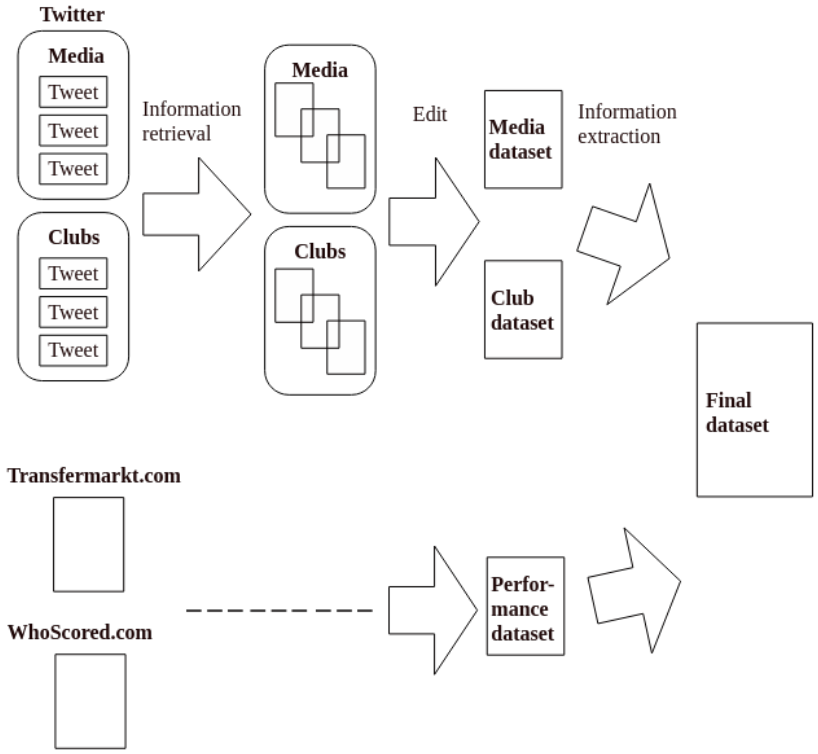

Figure 2. Text mining framework.

Information retrieval.

Previously, it was mentioned that the sample consisted of EPL teams. Table 1 lists the twenty teams used for the analysis, as well as a small description of the main extracted values. It can be noted that Arsenal has more than 9 million number of followers, the highest value of our sample. The team that tweeted the most in was Manchester City, with 12 540 tweets between August 2015 and May 2016.

Table 1.

Description of clubs accounts.

Club	Twitter account	Number of	Number of

			followers	tweets
1	Bournemouth	@afcbournemouth	259 793	7 262
2	Arsenal	@Arsenal	9’316 725	12 081
3	Aston Villa	@AVFCOfficial	914 365	6 481
4	West Ham United	@WestHamUtd	1’073 728	13 938
5	Manchester United	@ManUtd	1’0367 156	8 128
6	Norwich City	@NorwichCityFC	464 969	7 248
7	Leicester	@LCFC	933 458	7 551
8	Manchester City	@ManCity	4’063 464	12 540
9	Tottenham Hotspur	@SpursOfficial	1’864 438	9 934
10	Chelsea	@ChelseaFC	8’101 524	9 762
11	Newcastle United	@NUFC	1’001 006	5 337
12	Liverpool	@LFC	6’951 363	8 122
13	Swansea	@SwansOfficial	688 928	7 648
14	Watford	@WatfordFC	299 259	5 735
15	Everton	@Everton	1’046 020	8 896
16	West Bromwich Albion	@WBA	581 512	8 947
17	Crystal Palace	@CPFC	481 507	10 925
18	Sunderland	@SunderlandAFC	660 866	7 606
19	Southampton	@SouthamptonFC	712 443	9 723

20	Stoke City	@stokecity	687 573	5 953

Table 2 lists the media organizations used to provide part of the external variables. The media organizations can be divided in three groups. The first one is newspapers. The second one is TV Broadcasters. The third one is online media. It should be noted that The Guardian, BBC Sport and Bleacher Report accounts' focus is not only in football, but on many sports. The motivation to choose the seven organizations was based on their popularity (number of followers) in Twitter and experience.

Table 2.

Description of media accounts.

	Media organization	Twitter account	Number of followers	Number of tweets	Date joined
1	BBC Sport	@BBCSport	6'578 939	35 007	March 2011
2	Guardian Sport	@guardian_sport	722 979	21 404	June 2009
3	Bleacher Report	@br_uk	244 203	12 922	March 2007
4	The Independent Football	@IndyFootball	69 058	11 901	March 2012
5	Mirror Daily Football	@MirrorFootball	434 248	37 788	October 2008
6	Sky Football	@SkyFootball	3'217 909	18 028	July 2002
7	The Sun Football	@TheSunFootball	339 351	17 405	February 2009

The information retrieval step took a considerable amount of time because of constraints established by Twitter that allow only 900 tweets to be retrieved in a 15-minute window, and the programming required to retrieve the information. Due to the restrictions, we needed to 'rehydrate' a tweet. In other words, to obtain all the data we were required to go through Twitter's Application Programming Interface (API). Ten fields were extracted from each tweet: creation date, user name, users' followers, number of retweets, number of favorites, language, Id. of the replied user (in case the tweet was a reply), the name of the replied user, Id. Number, and text. In summary, the information retrieval process was composed like this:

1. Identify all Twitter media and club accounts.
2. Extract the Id. Number of all tweets published by each account (media and club) between August 2015 and May 2016.
3. 'Rehydrate' the tweets of each account using the API from Twitter.
4. Create a specific .txt file for each account with the corresponding tweet information.
5. Merge all club accounts in one dataset and all media accounts in another dataset.
6. Clean and edit datasets. Safely remove errors created by the content of the tweet. (E.g. large tweets)

Text analysis.

The second step in the text mining framework is text analysis. From this exploration, it is easily seen that clubs have more tweets than the media organizations. The club dataset contains 170 754 tweets and the media dataset consists of 154 455 tweets. All the teams used their Twitter account to display news, promote fans campaigns and narrate each match. The text was usually published along with the multimedia content of the team or a specific player. All content was written in English. When comparing

the budget size of each team, it can be determined that big budget clubs have a higher number of followers and small budget clubs have a lower number of followers.

In the case of the media accounts, tweets retrieved contained links, images, video or just text about all clubs in the Premier League. All content was written in English. Like the clubs' accounts, most of the tweet activity developed during the matches. The season used for the sample, 2015/2016 presented a curious case. Leicester City, a small club, became the champion for the first time in the history of the Premier League. The title claim by the 'Foxes' was a shock for the football atmosphere around the world. Thus, most of the editorial decisions were between the surprising small team and what was happening to the 'big' teams.

Information extraction.

While the collected data allow us to extract all kinds of different information, this research focuses on the count of players' Twitter accounts and their complete name. To search for the Twitter accounts, we rely on the BBC Sports Twitter list named Premier League and on tweetsfc.com, a website focused on gathering information from football Twitter accounts. After linking the corresponding user names with the full names in our datasets, we search manually for the players who appeared to have no Twitter account. In order to obtain a reliable measure and not leave out any player mention, we also took into account the full name of the player. Consecutively, we created a Python coding block to search and add the number of times each player was mentioned in media and club tweets, respectively. Also, the number of retweets of each tweet was added as a new variable. A retweet is a sharing option that implies the message is published to all the followers of the user who retweeted ('FAQs About Twitter', 2017).

3.2 Dataset

The dataset used for the regression models contained 15 performance variables and 9 external variables. The performance parameters (See Figure A.1) were obtained from Transfermarkt.com and WhoScored.com. The first one is a crowdsourcing platform that estimates the market value of players based on their individual performance weighted with the opinion of several parties, like journalists, fans, coaches, agents, and experts from different areas. The second website's main aim is to rate the performance of teams and players based on statistics updated daily. WhoScored.com consists of a team of analysts and software developers with a background in the sports industry. To avoid any change in the data, this research is based on a previous season to 2016/2017.

Table 3 lists and describes all the variables used for the analysis and Table 4 contains descriptive statistics for each variable.

Table 3.

Description of variables

	Variable	*Description*	*Type of variable*
X_1	Weight	Weight of the player in kilograms.	Performance
X_2	Appearances	Number of games the player started.	Performance
X_3	Substitutions	Number of games the player entered as a substitute.	Performance
X_4	Goals	Total number of goals scored in the season 2015/2016	Performance
X_5	Assists	Number of assists the player created in the season 2015/2016.	Performance
X_6	Yellow cards	Number of yellow cards in all the season.	Performance
X_7	Red cards	Number of red cards in all the season.	Performance
X_8	Shots	Average number of Shots to the goal per game.	Performance
X_9	Passes	Average percentage of successful passes.	Performance

X_{10}	Aerials	Aerials duels won per game.	Performance
X_{11}	Position	The position on the field (1-13)	Performance
X_{12}	Age	Age of the player at the end of the season.	Performance
X_{13}	Height	Height of the player in centimeters.	Performance
X_{14}	Foot	Striking foot. Left, right or not defined. (0-2)	Performance
X_{15}	Minutes	Total number of minutes played in the season.	Performance
X_{16}	Man of the Match	Number of times the player was selected by WhoScored.com as the best player of the match.	External
X_{17}	Followers	Number of followers in Twitter.	External
X_{18}	Mentions	Sum of mentions of the player in media tweets.	External
X_{19}	Retweets	Sum of retweets in media tweets the player was mentioned.	External
X_{20}	Mentions by Clubs	Sum of mentions of the player in club tweets.	External
X_{21}	Retweets by Clubs	Sum of retweets in club tweets the player was mentioned.	External
X_{22}	Clubs mentions	Number of times the name of a club was mentioned in media tweets.	External
X_{23}	Broadcast Mentions by Media	Sum of the mentions of the player multiplied by the number of followers each media account had.	External
X_{24}	Broadcast Mentions by Clubs	Sum of the mentions of the player multiplied by the number of followers each club account had.	External
X_{25}	Goals * Shots	A player who scores needs to shoot, but not all players who shoot score goals.	Interaction
X_{26}	Goals * Position	Forwards and attacking midfielder are expected to score more goals than players in other position.	Interaction
X_{27}	Goals * Followers	Every player that scores has a different level of exposure in social media.	Interaction
Y_1	Market value	The value (€) of a player, according to transfermarkt.com.	Dependent

Table 4.

Summary of the data by variables

	Variable	*Mean/SD*	*Max.*	*Min.*	*IQR*
X_1	Weight (kg.)	75.79/7.32	98 (R. Elliot)	58 (I. Anya)	11
X_2	Appearances	21.21/10.45	38 (9 players)	0 (75 pl.)	17
X_3	Substitutions	4.92/4.86	22 (L. Ulloa)	0 (134 pl.)	7
X_4	Goals	4.04/4.43	25 (H. Kane)	0 (294 pl.)	4
X_5	Assists	2.25 /2.85	19 (M. Ozil)	0 (304 pl.)	3
X_6	Yellow cards	2.98/2.32	11 (J. Colback)	0 (93 pl.)	3
X_7	Red cards	0.13/0.42	3 (V. Wanyama)	0 (507 pl.)	0
X_8	Shots	1.24/0.83	4.3 (P. Coutinho)	0 (105 pl.)	1.1
X_9	Passes	78.67/6.91	100.0 (15 pl.)	0 (5 pl.)	9
X_{10}	Aerials	1.28/1.17	6.5 (R. Gestede)	0 (85 pl.)	1.6
X_{11}	Position	-	-	-	
X_{12}	Age	25.32/3.29	39 (S. Given)	16 (5 pl.)	5
X_{13}	Height (cm.)	181.41/6.47	202 (C. Pantilimon)	165 (3 pl.)	9
X_{14}	Foot	-	-	-	-
X_{15}	Minutes	1895/889.69	3420 (4 pl.)	1 (3 pl.)	1482
X_{16}	Man of the Match	0.90/1.69	10 (R. Mahrez)	0 (366 pl.)	2
X_{17}	Followers	835211/ 2166288	14463871 (W. Rooney)	132 (M. Wasilewski)*	610678
X_{18}	Mentions	95/167	1349 (W. Rooney)	0 (63 pl.)	89
X_{19}	Retweets	2993 /5428	39394 (W. Rooney)	0 (63 pl.)	2699
X_{20}	Mentions by Clubs	157/141	835 (H. Kane)	0 (5 pl.)	138

X_{21}	Retweets by Clubs	34218/68641	406194 (A. Martial)	0 (5 pl.)	28415
X_{22}	Clubs mentions	42745/123	118167 (Leicester)	5587 (Norwich)	47010
X_{23}	Broadcast Mentions by Media	109529433 /183730765	1532435123 (W. Rooney)	0 (63 pl.)	86217764
X_{24}	Broadcast Mentions by Clubs	544760837/ 916256263	5380765861 (M. Ozil)	0 (63 pl.)	448622666
Y_1	Market value	11612139/ 9963460	60000000 (E. Hazard)	50 000 (K. Stewart)	12500000

*= *Only players with Twitter accounts were considered.*

The dataset comprised different information regarding each player. Here are some of the overall key factors:

- 559 total number of players.
- 365 players have a Twitter account.
- Media organizations did not mention 230 of the 365 players' Twitter accounts.
- Media organizations did not mention 63 footballers by their full name.
- 41 players had no corresponding market value because they didn't play any match or were too young to gather any type of information about them.

3.3 Linear regression

Most studies in this area use an econometric model for predicting transfer fees or market values (e.g. Carmichael, 2006; Herm et al., 2013; Franck & Nuesch, 2012). A linear regression estimates the value of one dependent variable based on independent variables (Tompkins, 1992). After the text mining process, the data allows us to detect patterns and to build a regression model. In previous research (e.g. Cody et al., 2016; Celli et al., 2016), words have been used as variables to predict behavior. For this

research, the model will be composed of performance (talent of a player) and external (data related to the media and social media) variables.

Interactions.

The nature of football establishes certain natural interactions between variables which can be transformed into statistical interactions. An interaction describes the effect of one independent variable over the dependent variable that may depend on the level of another independent variable (Jaccard & Turrisi, 2003). For this study, three interactions were taken into account. The first is the effect of Goals and Shots. Every player who scores shoots to target, but not all players who shoot score goals. The second involves Goals and Position. Forwards and attacking midfielders are expected to score more goals than players in other positions. The third interaction was between Goals and Followers. The effect of number of goals is not the same for players with high and low social media exposure.

Regression Evaluation.

Two main evaluation statistics are introduced in this study. The regression model that scores higher than the two others in the following two coefficients will be considered as the best option to predict the market value of a player in this study case.

Adjusted R-squared.

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. This statistic in mention is a modified version of R-squared that has been adjusted for the number of predictors in the model. The main objective is to reduce the effect of multiple variables on a predictive model. It decreases when a predictor improves the model by less than expected by chance. It is always lower than the R-squared (Steel & Torrie, 1960).

BIC

The second technique to compare models is the Bayesian Information Criterion (BIC). This statistical selection tool selects among a finite set of models the model with the lowest BIC (Schwarz, 1978). “The BIC is intended to provide a measure of the weight of evidence favoring one model over

another, or Bayes factor" (Weakliem, 2016). As opposed to the AIC, it is a measure that penalizes models for each extra variable included in the model. Therefore, BIC will pursue to find models with fewer variables (Edwards et al., 1963).

3.4 Estimating regression models

In order to test the influence of media in football influencers of transfermarkt.com, we established two main approaches. One with 'broadcast' measures, which includes two new variables: Broadcast Mentions by Media, and Broadcast Mentions by Clubs (see Table 3). The second approach includes parameters like mentions in media accounts and club accounts without any broadcasting effect. As stated before, the focus of the following models is in prediction and not in explanation. The foremost analysis is to understand correlations, which are not an indicator of cause and effect.

Variable transformation.

The distribution of the football market value along with transfers had shown to be an exponential statistical distribution. Most of the footballers had a similar value in the market, while few stand out and raise the average market value. The number of mentions and retweets extracted for our study present the same kind of distribution. Most have few or none mentions, while the 'superstars' of the bigger clubs are mentioned in a higher proportion. Therefore, the logarithm of base 10[1], which reduces the magnitude but does not influence variability (Benoit, 2011), was taken for the following variables: Clubs mentions, Mentions by clubs, Retweets in clubs, Followers, Retweets, Mentions, Broadcast Mentions by media and Broadcast Mentions by clubs. Position and Foot were transformed in dummy variables (See Annexes). Likewise, the performance variables and the Man of the Match variable were centered because many of them have different scales.

[1] We add 1 unit to all the values of the variables that were to be transformed to logarithm of base 10 because the $\log_{10}$ of 0 is infinity.

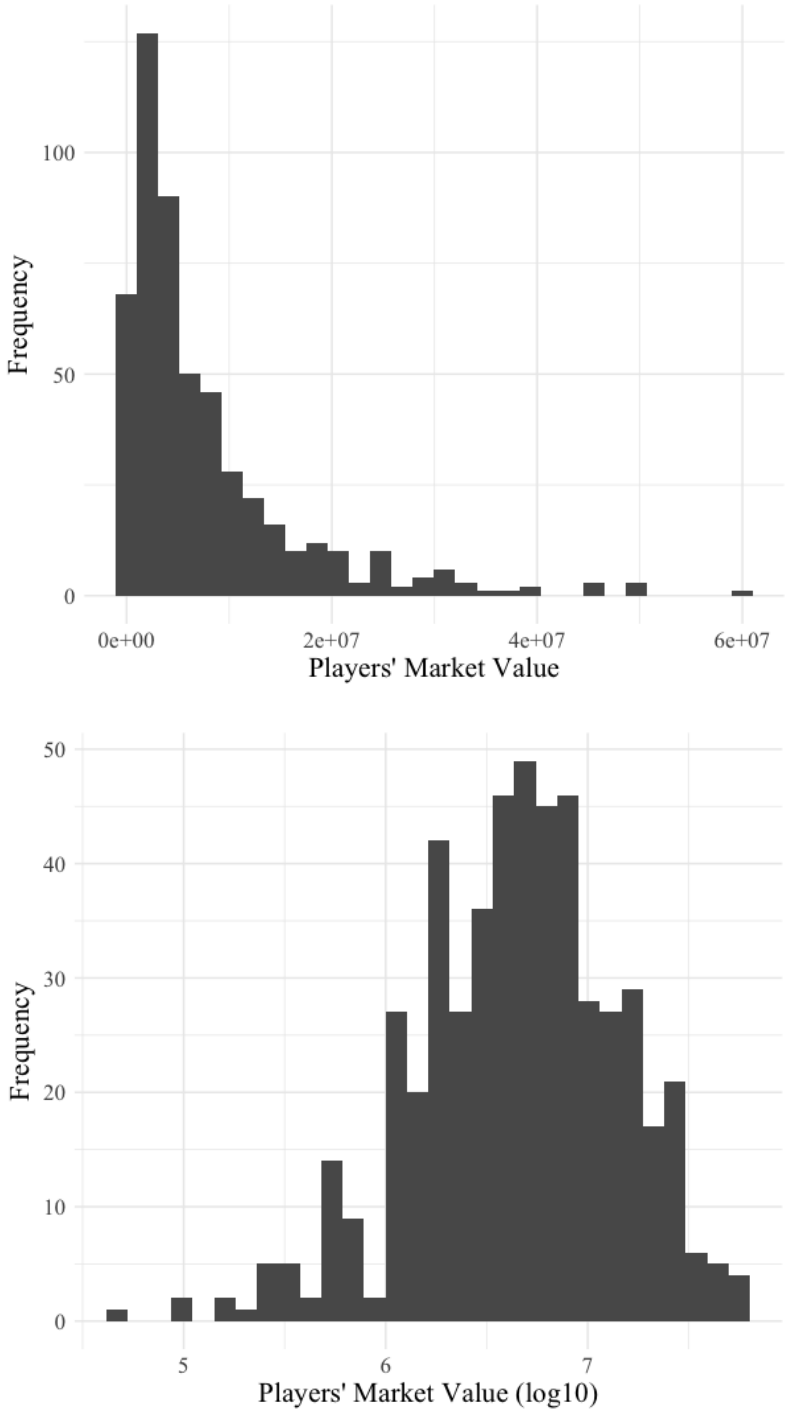

Figure 3. Histogram showing player`s market value without (above) and with (below) logarithm of base 10 transformation.

Correlation.

Before establishing a regression model, Pearson correlation coefficients between the employed variables in our analysis were calculated. The goal was to identify variables with a high correlation that could hurt the regression model performance. If one of the variables scored a higher correlation coefficient of 0.90 with another, the one with the lowest correlation value with the dependent variable was left out from the two models (broadcast effect and non-broadcast effect) composed for the analysis.

Table 5.

Highly correlated variables

Variables / Correlation	Variable removed / (correlation with Market value)
Minutes – Appearances = 0.996	Minutes (0.297)
Mentions – Retweets = 0.937	Retweets (0.468)

Feature selection.

We used an automated and efficient approach for choosing a smaller set of models to consider. By using a function of subset selection from an R package named "leaps", that "performs an exhaustive search for the best subsets of the variables in x for predicting y in linear regression, using an efficient branch-and-bound algorithm" (Lumley, 2017). The outcome presented different model combinations. Starting with the two full models, including 21 (non-broadcast effect) and 19 variables (broadcast effect), respectively, and ending with just one independent predictor. Aside from the broadcast and non-broadcast models, we built a third model including only performance parameters.

The objective of this study is to obtain three regression models in total. One that shows patterns with broadcast effect variables, another without the broadcast effect, and a third with only performance variables. Each approach will have the best subset model based accordingly to the BIC. After, the Adjusted-R^2 will provide a statistical background to compare and interpret the three models.

The performance model:

$$\log_{10}(Y_1) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14}$$

The non-broadcast model:

$$\log_{10}(Y_1) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{16} X_{16} + \beta_{17} \log_{10}(X_{17}) + \beta_{18} \log_{10}(X_{18}) + \beta_{20} \log_{10}(X_{20}) + \beta_{21} \log_{10}(X_{21}) + \beta_{22} \log_{10}(X_{22})$$

The broadcast model:

$$\log_{10}(Y_1) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{16} X_{16} + \beta_{17} \log_{10}(X_{17}) + \beta_{22} \log_{10}(X_{22}) + \beta_{23} \log_{10}(X_{23}) + \beta_{24} \log_{10}(X_{24})$$

Apart from the three full base models described above, our analysis takes into account three interactions. This means we run the same broadcast model plus the interactions and the non-broadcast model plus the interactions. Neither of the three new parameters had a significant effect on the response variable, nor reached as part of the best subset selection (See Table A.1). The outcome was identical to the ones presented in the next section.

4. Results

Table 6 shows the results of the regression-based analysis on the market values of 173 Premier League football players in the season 2015/2016. Table A.2 (appendix) comprises the correlation between all variables used in this study. As explained before, the aim of our study is to compare how different is the relationship between weighted (broadcast effect) and non-weighted (non-broadcast effect) responses, so we built two models. Also, we analyze the two models including interactions but the subset selection feature considered them non-significant. Each model was tested for heteroscedasticity and multicollinearity, using the Breush-Pagan method and the Variance Inflation Factor (VIF), respectively.

Table 6.

Regression models

Dependent variable: Market value			
Independent variable	Model 1 *(Performance)*	Model 2 *(No broadcast effect)*	Model 3 *(Broadcast effect)*
Weight	0.001		
	(0.007)		
Apps	-0.004		
	(0.005)		
Substitutions	-0.008		
	(0.009)		
Goals	-0.017	-0.030***	-0.030***
	(0.014)	(0.010)	(0.010)
Assists	0.045***		
	(0.014)		
Yellow cards	0.018	0.036***	0.032***
	(0.017)	(0.012)	(0.011)
Red cards	0.015		

	(0.079)		
Shots	0.268***	0.258***	0.232***
	(0.071)	(0.053)	(0.053)
Passes	0.023***		
	(0.006)		
Aerials	0.038		
	(0.045)		
Position	0.004		
	(0.011)		
Age	0.016		
	(0.011)		
Height	0.002		
	(0.009)		
Foot	-0.110		
	(0.077)		
Followers		0.375***	0.289***
		(0.036)	(0.048)
Broadcast mentions by clubs			0.173***
			(0.066)
Constant	6.661***	4.778***	3.805***
	(0.095)	(0.188)	(0.415)
Observations	173	173	173
R^2	0.296	0.491	0.512
Adjusted R^2	0.234	0.479	0.497
Residual St. Error	0.423 (df = 158)	0.349 (df = 168)	0.343 (df = 167)
F- Statistic	4.746*** (df = 14; 158)	40.593*** (df = 4; 168)	34.973*** (df = 5; 167)
BIC	16.571619	-91.22610	-93.02129

*Note: *p<0.1; **p<0.05; ***p<0.01*

Model 1 includes only the performance parameters extracted from Transfermarket.com and WhoScored.com. We assume that player skills, following Herm et al., (2013) study, are the base to build a criterion around the price value of a footballer. The linear regression proposed for Model 1 showed Shots, Passes, and Assists have a significant influence ($p<0.01$) on market value. The interpretation is that the more a player shoots to goal, the higher his market value. A unit increase on the average of shots per game will result approximately in a 27% increase in the market value. The same logic with the average number of successful passes. When a player increases his percentage of pass accuracy the market value will result in a 2.4 % change. Also, if a footballer adds one more assist, a change of 4 % will occur in the predicted variable. Furthermore, the data don't replicate some of the findings in past studies (Bryson et al., 2013).

When entering the external variables into the model we found a vanishing influence of some of the players' performance variables highlighted before. Model 2, with non-broadcast effect parameters, is the result of the best subset selection technique according to the BIC. Three performance variables are significant. However, Goals presents a negative relationship with the predicted variable. According to the Goals coefficient, one more goal scored results in 3% decrease change in the market value, which is counterintuitive with reality. Yellow cards and Shots have a positive relationship with market value. An additional yellow card in the season would transform into 3.6 % change in the predicted parameter. Similarly to what the Model 1 showed, Shots is the leading change variable. In Model 2, an increase of a unit in the average shots per game results in a 25.8% change in the market value. The fourth variable that completes Model 2, is Followers. The number of users that click on the follow button of each players' account has a positive relationship with the dependent variable. One percentage increase in the number of followers results in a 37.5% change in the market value. Surprisingly, no other external variables

(Mentions, Mentions by Clubs, Retweets by Clubs, Man of the match or Clubs mentions) get to be chosen by the subset selection tool. In contrast to Herm et al (2013) study, the opinion of experts (Man of the match parameter) did not influence a footballer's market value.

Finally, we took the full model and made a subset selection including performance variables, and broadcast variables. Model 3 is the outcome of the subset selection based on the BIC with broadcast effect. In this model, Mentions, Retweets, Mentions by clubs, and Retweets by clubs are removed. Similarly to Model 2, Goals, Yellow cards and Shots have a significant relationship with the predicted variable. Goals keep showing a counter-intuitive coefficient (-3%). While Yellow cards and Shots have positive effects (3.2% and 23.2%, respectively) in the change of the market value. The number of followers repeats as an influential parameter. In Model 3, a 1% increase in the number of followers transforms in a 28.9% change in the market value. Finally, the Broadcast mentions by clubs parameters show a significant positive relationship with the controlled variable. An increase of 1% in this value results in 17.3% change in the market value.

In terms of the Adjusted-R^2, Model 1 scored a 0.23 of explanatory power, the lowest of all three models presented in this study. When comparing between Model 2 and Model 3 there is a small difference. The former has an Adjusted-R^2 of 0.48, while the latter scored a 0.50. As explained before, the Adjusted-R^2 increases only if the new parameter improves the model more than it would be expected by chance. As we can see, all performance variables don't perform better than a specific subset. Also, a complete combination of performance and external variables doesn't transform into a better predictive model. In this case, Model 3, with broadcast effect variables, is the best option to predict the players' market value.

The BIC is the measure that stated which subset model was the better option for the two models. The BIC for Model 1, shown in Table 6, was calculated using the same process than the other two but the coefficient (16.57) corresponds to the complete model, the highest of all the three models. Model 2 and

Model 3 come from two different base models. Both models included the same number of performance variables but differ in the external variables. For the model with non-broadcast measures, the subset with four parameters scored the lowest BIC (-91.23), and for the model with broadcast effect, the subset of five parameters was chosen (BIC = -93.02).

In the three models, the Variance Inflation Factors (VIF) indicated no problems of multicollinearity (Table A.4). According to Hutcheson and Sofroniou (1999), the VIF statistic that exceeds 5 indicates a problem with multicollinearity. Finally, we test for heteroscedasticity using the Breusch-Pagan test (See Table A.5). In simpler terms, this means that the variance of residuals should not increase with fitted values of the response variable (Prabhakaran, 2016). All three models showed higher p-values than 0.01, which is considered the threshold to determine if a model suffers from heteroscedasticity.

5. Discussion

This research investigates what influences football enthusiasts when they estimate a player's market value. Following the study of Herm et al. (2013), we add a new case study of how an online community performs well on the complex task of human capital evaluation. The direction of this study emphasizes the role of media in football transfers. Accordingly, this research contributes to previous literature a comparison between weighted and non-weighted word frequency variables. As shown in the results section, one model performed slightly better with a weighted parameter (Broadcast mention by clubs) than with regular frequency count values.

The analysis described in previous sections is an online adaptation of the Brunswik's lens model. We replicate the model using performance and external variables. Although each crowdsourcing participant has his own criteria to rate a player, our analysis results indicate which variables have a stronger effect on them. Also, the broadcast effect model provides a reflection of the sources that influences football enthusiasts. The role of the media in this case was not significant in comparison with the work of Public Relations (PR) and marketing departments, which control the club's Twitter accounts. Price et al. (2013) introduced and promoted in their study the referee journalist role, a reporter who offers an informed opinion against the white noise of Twitter. However, the variable with a broadcast effect in clubs has a stronger relationship with the response variable. Hence, PR and marketing in an emotional environment like football, relate better than the media.

5.1 Limitations

The constraints of this research start with the data sample. We use information from one of the more important leagues in the world, the English Premier League, but we only picked data from one season (2015/2016). Also, we chose only seven, well-recognized British media organizations, discarding small or more specialized companies. Likewise, choosing only one social network (Twitter) leaves out all

the discussion created through other channels. However, we believe that taking data from one social network and combining the three main actors (media, players, and clubs) was a good reflection of reality.

Other limitations of the data were the ambiguity of certain words like 'Liverpool' and 'Leicester'. Both words could be used to describe a city or the football clubs. To diminish the possible effect of counting mistaken words, we made several decisions. First, we chose seven media accounts specialized in sports, so the chance of mentioning an event happening in both cities decreased. Second, when extracting the word count, we included the Twitter username and the full name of both clubs: 'Liverpool Football Club' and 'Leicester City Football Club'. One more constraint was that Twitter does not allow retrieving the exact number of followers that an account had when a specific tweet was published. Hence, the number of followers is the number when data was retrieved. Finally, recent literature has demonstrated different additional variables (e.g. mental abilities, speed, etc.) could affect player selection.

5.2 Further research

Taking the limits of this study into consideration, we strongly suggest future studies to conduct research using text-mining techniques for regression tasks. Also, the present research could be used as an example to use alternative econometric models to approximate actual responses in online communities. Furthermore, as mentioned by Herm et al. (2013), a deeper analysis of the content that is discussed in crowdsourcing communities would be better for comparisons and evaluations.

Another challenge for future investigations around this topic is the use of different econometric approaches. The feature selection could also be determined with different techniques. In terms involving the analysis itself, we suggest using more specialized media companies, like magazines, television shows or websites only dedicated to football. Similarly, football enthusiasts may trust specific journalists as sources to detect trends related to transfers rather than big Twitter media accounts. In addition, further research can include parameters referred to sports agents and managers influence.

6. Conclusion

The market value of a player is not the result of an exact formula. Our specific aim was to test the influence of media and social media in football enthusiasts that participate in a crowdsourcing task through a linear regression. We are able to conclude that some media and social media values have a strong relationship with the player's market value. Hence, they can be used as indicators to football enthusiasts, sports journalists or sports managers for under or overvalued players. After all the analysis, the research questions can be answered as follows:

RQ1: To what extent do external variables and performance variables relate to the player's market value?

Based on the regression analysis, we can conclude that the model with broadcast effect (Model 3, Adjusted-R^2 = 0.50) has a stronger relationship with our response variable than a model that included only performance variables (Model 1, Adjusted-R^2= 0.23) and one with non-broadcasting effect (Model 2, Adjusted-R^2=0.48). Model 3 is composed of three performance parameters and only two external variables, which are strongly significant ($p<0.01$) when predicting the market value of a player. Goals, yellow cards, and shots proved to be crucial performance features to be considered by participants of a crowdsourcing task. Also, Followers and Broadcast mentions by clubs completed the variables that built Model 3. The broadcast weight of the clubs accounts performed better than mentions in media.

RQ2: Is the size of the audience measured by number of followers, a proxy of a broadcast effect, a suitable predictor to forecast the player's market value?

A component of Model 3 is Broadcast mentions by clubs. The other broadcast variable, named Broadcast mentions by media, was not chosen in the best subset selection process. Therefore, the audience of all the 20 clubs can be considered a suitable predictor of the response variable. In the case of the media, the audience of the seven organizations was not a crucial factor. Beyond the regression coefficients, this result means that the social media discussion of all the teams could be more precious for

football aficionados. First, the interaction between clubs in social media increases with time. But, the main purpose of those accounts is to promote a good image of their own. Furthermore, when one team mentions another team's player, the value of that mention is greater than a media organization mention. The audience, number of followers, of each team is a reflection of the number of fans it attracts. The recognition of rivals, we must say, gives an interesting weight of the player's mentions. In conclusion, assigning a weight to a unit proves to be suitable when the parties involved might have different interests (e.g. each account tries to promote their own team, not the opposition team) but a common truth (e.g. a goal scored by a player of the opposition team).

7. References

Antonioni, P., & Cubbin, J. (2000). The Bosman ruling and the emergence of a single market in soccer talent. *European Journal of Law and Economics, 9(2),* 157-173.

Arceo, A. (2003). Beckham un fenómeno de mercado. *Chasqui, 83,* 4-11.

Adler, M. (1985). Stardom and talent. *American Economic Review, 75,* 208–212.

Benoit, K. (2016). Linear regression models with logarithmic transformations. *London School of Economics, 1,* 1-8. Retrieved from http://www.kenbenoit.net/courses/ME104/logmodels2.pdf

Barajas, A., Sánchez, P., & Urrutia de Hoyos, I. (2010). El Mercado de traspasos de futbolistas: un análisis internacional. *Revista Decisión, 10*, 31-55.

Bryson, A., Rossi, G., & Simmons, R. (2014). The migrant wage premium in professional football: a superstar effect? Doi: 10.1111/kykl.12041

Cataldi, M., & Aufaure, M. (2014). The 10 million follower fallacy: audience size does not prove domain-influence on Twitter. *Knowledge and Information Systems, 44(3),* 559-580. http://dx.doi.org/10.1007/s10115-014-0773-8

Carmichael, F. (2006). The player transfer system in soccer. *Handbook on the Economics of Sport, 1,* 668-676. Retrieved from http://www.ahmetguvener.com/wp-content/uploads/Handbook-on-the-Economics-of-Sport.pdf

Carmichael, F., Forrest, D., & Simmons, R. (1999). The labour market in association football: who gets transferred and for how much? *Bulletin of Economic Research, 2*, 125–150.

Celli, F., Stepanov, E., Poesio, M., & Riccardi, G. (2016). Predicting Brexit: classifying agreement is better than sentiment and pollsters. *School for Computer Science and Electronic Engineering,* 110-118. University of Essex, UK. Retrieved from https://goo.gl/8Gg9tP

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. *ICWSM, 10,* 10-17. Retrieved from http://twitter.mpi-sws.org/icwsm2010_fallacy.pdf

Charness, G., & Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives, 26*, 157–176.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci., 10,* 103-126. Retrieved from http://www.annualreviews.org/doi/pdf/10.1146/annurev.polisci.10.072805.103054

Cody, E., Reagan, A., Dodds, P., & Danforth, C. (2016). Public opinion polling with Twitter. *Department of Mathematics & Statistics,* 1-15. University of Vermont. Retrieved from https://arxiv.org/abs/1608.02024v1

Cody, E., Reagan, A., Mitchell, L., Dodds, P., & Danforth, C. (2015). Climate change sentiment on Twitter: an unsolicited public opinion poll. *Plos One, 10(8)*, e0136092. http://dx.doi.org/10.1371/journal.pone.0136092

Dobson, S., & Gerrard, B. (1999). The determination of player transfer fees in English professional soccer. *Journal of Sport Management, 13(4),* 259-279. http://dx.doi.org/10.1123/jsm.13.4.259

Dombert, S. (2017). *Special Request* [E-mail].

De Vreese, C. H., Boomgaarden, H. G., & Semetko, H. A. (2011). (In) direct framing effects: the effects of news media framing on public support for Turkish membership in the European Union. *Communication Research, 38(2),* 179-205.

Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70(3),* 193-242. http://dx.doi.org/10.1037/h0044139

Fahy, D., & Nisbet, M. (2011). The science journalist online: shifting roles and emerging practices. *Journalism, 12(7),* 778-793. http://dx.doi.org/10.1177/1464884911412697

FAQs about Twitter. (2017). *Twitter Help Center.* Retrieved 6 July 2017, from https://support.twitter.com/articles/77606

Franck, E., & Nüesch, S. (2010). Talent and/or popularity: what does it take to be a superstar? *Economic Inquiry, 50(1),* 202-216. http://dx.doi.org/10.1111/j.1465-7295.2010.00360.x

Frank, R. H. y Cook, P. J. (1995). *The winner-take-all society.* New York: Martin Kessler Books.

Foster, R. (2016). *How football clubs calculate the cost of buying players in the transfer market.* The Guardian. Retrieved 7 February 2017, from https://goo.gl/pLaTZ8

Galton, F. (1907). Vox populi. *Nature,* 450.

Gans, H. J. (1979). *Deciding what's news: a study of CBS evening news, NBC nightly news, Newsweek, and Time.* Northwestern University Press.

Garcia del Barrio, P., & Pujol, F. (2007). El papel del fútbol en la sociedad actual. *Fútbol: ocio y negocio. Revista Empresa Y Humanismo, 11,* 89-108.

Gerhards, J., Mutz, M., & Wagner, G. (2014). Predictable winners. Market value, inequality, diversity, and routine as predictors of success in European soccer leagues. *Zeitschrift für Soziologie, 43,* 483501.

Gitlin, T. (1980). *The whole world is watching: Mass media in the making & unmaking of the new left.* University of California, Press.

Goldstein, N. (2017). *The why and when of centering continuous predictors in regression*. Retrieved 7 July 2017, from https://goo.gl/3HiKPq

Haynes, R. (2007). Footballers' Image Rights in the New Media Age. *European Sport Management Quarterly. 7(4),* 361-374. Retrieved from https://www.stir.ac.uk/research/hub/publication/834

Herberger, T., & Wedlich, F. (2016). What athletic characteristics determine professional football players' market values: a crowdsourced valuation. *Discussion paper series in economics and management, 16-25.*

Herm, S., Callsen-Bracker, H., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review, 17(4),* 484-492. http://dx.doi.org/10.1016/j.smr.2013.12.006

Hutcheson, G., & Sofroniou, N. (1999). *The Multivariate Social Scientist*. Sage Publications Ltd. London, UK.

Hutchins, B. (2011). The acceleration of media sport culture. *Information, Communication & Society, 14(2)*, 237-257. http://dx.doi.org/10.1080/1369118x.2010.508534

Jaccard, J. & Turrisi, R. (2003). *Interaction effects in multiple regression.* Sage University Papers Series on Quantitative Applications in the Social Sciences. Thousands Oaks, US.

Lippmann, W. (1946). *Public opinion*. Transaction Publishers. New Brunswick, US.

Lumley, T. (2017). *Package 'leaps'* [PDF guide]. CRAN. Retrieved from https://cran.r-project.org/web/packages/leaps/leaps.pdf

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly, 36(2)*, 176-187.

Mellado, C. (2014). Professional Roles in News Content. *Journalism Studies, 16(4),* 596-614. http://dx.doi.org/10.1080/1461670x.2014.922276

Prabhakaran, S. (2017). *How to detect heteroscedasticity and rectify it?* DataScience. Retrieved 1 July 2017, from https://datascienceplus.com/how-to-detect-heteroscedasticity-and-rectify-it/

Price, J., Farrington, N., & Hall, L. (2013). Changing the game? The impact of Twitter on relationships between football clubs, supporters and the sports media. *Soccer & Society, 14(4),* 446-461. http://dx.doi.org/10.1080/14660970.2013.810431

Rottenberg, S. (1956). The Baseball Players' Labor Market. *Journal of Political Economy, 64(3),* 242-258. http://dx.doi.org/10.1086/257790

Rosen, S. (1981). The Economics of Superstars. *Journal of Political Economy, 79,* 1302-1319.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6 (2),*

461–464. doi:10.1214/aos/1176344136.

Shoemaker, P. & Reese, S. (1996). *Mediating the message: theories of influences on mass media content.* Longman Publishers. New York, US.

Shmueli, G. (2010). To Explain or to predict? *SSRN Electronic Journal.* http://dx.doi.org/10.2139/ssrn.1351252

Speight, A. and Thomas, D. (1997). Arbitrator decision-making in the transfer market: an empirical analysis. *Scottish Journal of Political Economy, 44*, 198–215. doi:10.1111/1467-9485.00053

Steel, R. G. D. & Torrie, J. H. (1960). *Principles and Procedures of Statistics with Special Reference to the*

Biological Sciences. McGraw Hill.

Surowiecki, J. (2005). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Little Brown. UK.

Tan, A. n.d. *Text mining: the state of the art and the challenges.* Retrieved 3 July 2017 from https://pdfs.semanticscholar.org/9a80/ec16880ae43dc20c792ea3734862d85ba4d7.pdf

Tompkins, C. A. (1992). Using and interpreting linear regression and correlation analyses: Some cautions and considerations. *Clinical Aphasiology, 21,* 35-46.

Weakliem, D. L. 1999. A Critique of the Bayesian Information Criterion for Model Selection. *Sociological Methods & Research, 27*, 359-397.

Weaver, D., & Elliott, S. N. (1985). Who sets the agenda for the media? A study of local agenda-building. *Journalism Quarterly, 62(1),* 87-94.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives, American Economic Association, 18,* 107–126.

Appendix

Table A.1

Bayesian Information Criterion

Number of variables	*With interactions*		*Without interactions*	
	Non-broadcast	Broadcast	Non-broadcast	Broadcast
1	-75.29	-75.29	-75.29	-75.29
2	-84.27	-84.27	-84.27	-84.27
3	-87.07	-87.44	-87.07	-87.44
4	**-91.23**	-91.23	**-91.23**	-91.22
5	-89.21	**-93.02**	-89.20	**-93.02**
6	-87.08	-91.56	-86.78	-91.42
7	-84.40	-89.27	-84.39	-89.27
8	-82.25	-86.75	-82.25	-86.36
9	-79.61	-83.64	-79.61	-83.05
10	-77.93	-79.71	-77.93	-79.71
11	-73.74	-75.56	-73.74	-75.56
12	-69.82	-71.51	-69.82	-71.51
13	-65.21	-67.38	-65.21	-67.38
14	-60.42	-62.61	-60.42	-62.61
15	-55.52	-57.84	-55.52	-57.85
16	-50.60	-52.97	-50.60	-52.88
17	-45.69	-48.03	-45.69	-47.82
18	-40.60	-42.98	-40.58	-42.73
19	-35.50	-37.93	-35.47	-37.59
20	-30.39	-32.88	-30.34	
21	-25.28	-27.76		
22	-20.16	-22.61		
23	-15.01			

Table A.2.

Correlation Matrix

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27
X1	-0.028	1	0.172	-0.161	0.182	-0.023	-0.091	0.067	0.05	-0.074	-0.3	0.376	0.043	-0.13	0.258	0.728	-0.047	0.051	-0.059	0.048	0.005	-0.03	-0.055	-0.004	-0.072	-0.065	0.071	-0.012
X2	0.297	0.172	1	-0.209	0.996	0.422	0.454	0.644	0.218	0.314	0.055	0.303	0.542	-0.022	0.254	0.075	0.063	0.357	-0.007	0.363	0.431	0.266	0.287	0.319	0.346	0.148	0.329	0.27
X3	0.081	-0.161	-0.209	1	-0.167	0.136	0.056	-0.097	-0.032	0.212	0.103	-0.035	-0.112	0.011	-0.035	-0.166	-0.102	0.119	0.001	0.136	0.126	0.085	0.072	0.039	0.138	-0.106	0.148	0.11
X4	0.297	0.182	0.996	-0.167	1	0.424	0.453	0.644	0.216	0.314	0.048	0.32	0.538	-0.021	0.256	0.086	0.06	0.363	-0.007	0.37	0.437	0.268	0.286	0.318	0.343	0.134	0.338	0.274
X5	0.291	-0.023	0.422	0.136	0.424	1	0.522	0.205	0.088	0.784	0.051	0.177	0.642	-0.042	-0.013	-0.09	-0.013	0.483	0.084	0.408	0.428	0.359	0.267	0.93	0.823	0.447	0.295	0.276
X6	0.317	-0.091	0.454	0.056	0.453	0.522	1	0.262	0.066	0.507	0.112	0.019	0.518	0.024	0.008	-0.196	0.111	0.344	0.11	0.3	0.346	0.286	0.29	0.406	0.459	0.228	0.216	0.254
X7	0.192	0.067	0.644	-0.097	0.644	0.205	0.262	1	0.171	0.199	0.153	0.296	0.302	-0.057	0.167	0.02	0.011	0.19	-0.04	0.209	0.25	0.129	0.148	0.137	0.125	-0.026	0.215	0.156
X8	0.077	0.05	0.218	-0.032	0.216	0.088	0.066	0.171	1	0.081	0.036	0.125	0.035	-0.012	0.087	0.057	-0.021	0.099	0.01	0.106	0.035	0.036	0.02	0.067	0.068	-0.038	0.105	0.047
X9	0.395	-0.074	0.314	0.212	0.314	0.784	0.507	0.199	0.081	1	0.159	0.135	0.509	-0.026	-0.044	-0.183	-0.011	0.484	0.056	0.414	0.424	0.372	0.265	0.732	0.646	0.368	0.308	0.303
X10	0.22	-0.3	0.055	0.103	0.048	0.051	0.112	0.153	0.036	0.159	1	-0.003	0.056	-0.051	-0.036	-0.36	0.061	0.132	0.081	0.139	0.06	0.107	0.204	0.013	0.042	0.114	0.106	0.049
X11	0.072	0.376	0.303	-0.035	0.32	0.177	0.019	0.296	0.125	0.135	-0.003	1	0.184	-0.184	0.216	0.392	-0.016	0.135	-0.053	0.148	0.116	0.069	-0.004	0.13	0.074	-0.108	0.15	0.091
X12	0.262	0.043	0.542	-0.112	0.538	0.642	0.518	0.302	0.035	0.509	0.056	0.184	1	-0.043	0.04	-0.02	0.033	0.391	0.074	0.347	0.393	0.317	0.254	0.595	0.583	0.327	0.242	0.252
X13	-0.018	-0.13	-0.022	0.011	-0.021	-0.042	0.024	-0.057	-0.012	-0.026	-0.051	-0.184	-0.043	1	-0.092	-0.207	0.095	-0.008	0.009	0.039	-0.003	-0.003	0.042	-0.047	0.258	-0.016	0.047	0.001
X14	-0.02	0.258	0.254	-0.035	0.256	-0.013	0.008	0.167	0.087	-0.044	-0.036	0.216	0.04	-0.092	1	0.139	-0.07	0.079	-0.205	0.117	0.043	-0.061	0.211	-0.024	-0.042	0.08	0.204	-0.028
X15	-0.066	0.728	0.075	-0.166	0.086	-0.09	-0.196	0.02	0.057	-0.183	-0.36	0.392	-0.02	-0.207	0.139	1	-0.055	0.005	-0.069	-0.003	-0.069	-0.037	-0.116	-0.058	-0.152	-0.163	0.014	-0.034
X16	0.016	-0.047	0.063	-0.102	0.06	-0.013	0.111	0.011	-0.021	-0.011	0.061	-0.016	0.033	0.095	-0.07	-0.055	1	0.039	-0.016	0.036	-0.005	0.014	0.073	-0.016	0.024	0.004	0.031	-0.022
X17	0.536	0.051	0.357	0.119	0.363	0.483	0.344	0.19	0.099	0.484	0.132	0.135	0.391	-0.008	0.079	0.005	0.039	1	0.346	0.941	0.609	0.757	0.618	0.407	0.41	0.332	0.778	0.569
X18	0.252	-0.059	-0.007	0.001	-0.007	0.084	0.11	-0.04	0.01	0.056	0.081	-0.053	0.074	0.009	-0.205	-0.069	-0.016	0.346	1	0.306	0.164	0.485	0.343	0.09	0.071	0.113	0.176	0.318
X19	0.475	0.048	0.363	0.136	0.37	0.408	0.3	0.209	0.106	0.414	0.139	0.148	0.347	0.039	0.117	-0.003	0.036	0.941	0.306	1	0.545	0.692	0.557	0.327	0.354	0.306	0.892	0.52
X20	0.369	0.005	0.431	0.126	0.437	0.428	0.346	0.25	0.035	0.424	0.06	0.116	0.393	-0.003	0.043	-0.069	-0.005	0.609	0.164	0.545	1	0.738	0.474	0.362	0.348	0.275	0.447	0.747
X21	0.485	-0.03	0.266	0.085	0.268	0.359	0.286	0.129	0.036	0.372	0.107	0.069	0.317	-0.003	-0.061	-0.037	0.014	0.757	0.485	0.692	0.738	1	0.672	0.306	0.297	0.315	0.513	0.845
X22	0.609	-0.055	0.287	0.072	0.286	0.267	0.29	0.148	0.02	0.265	0.204	-0.004	0.254	0.042	0.211	-0.116	0.073	0.618	0.343	0.557	0.474	0.672	1	0.241	0.226	0.4	0.404	0.563
X23	0.251	-0.004	0.319	0.039	0.318	0.93	0.406	0.137	0.067	0.732	0.013	0.13	0.595	-0.047	-0.024	-0.058	-0.016	0.407	0.09	0.327	0.362	0.306	0.241	1	0.725	0.52	0.219	0.233
X24	0.224	-0.072	0.346	0.138	0.343	0.823	0.459	0.125	0.068	0.646	0.042	0.074	0.583	0.258	-0.042	-0.152	0.024	0.41	0.071	0.354	0.348	0.297	0.226	0.725	1	0.33	0.257	0.228
X25	0.298	-0.065	0.148	-0.106	0.134	0.447	0.228	-0.026	-0.038	0.368	0.114	-0.108	0.327	-0.016	0.08	-0.163	0.004	0.332	0.113	0.306	0.275	0.315	0.4	0.52	0.33	1	0.192	0.321
X26	0.304	0.071	0.329	0.148	0.338	0.295	0.216	0.215	0.105	0.308	0.106	0.15	0.242	0.047	0.204	0.014	0.031	0.778	0.176	0.892	0.447	0.513	0.404	0.219	0.257	0.192	1	0.412
X27	0.391	-0.012	0.27	0.11	0.274	0.276	0.254	0.156	0.047	0.303	0.049	0.091	0.252	0.001	-0.028	-0.034	-0.022	0.569	0.318	0.52	0.747	0.845	0.563	0.233	0.228	0.321	0.412	1

Table A.3

Heteroscedasticity

	Breusch-Pagan test	
Model	*BP*	*p-value*
Model 1	14.261	0.4304
Model 2	12.815	0.0122
Model 3	15.034	0.0102

Table A.4

Variance Inflation Factor (VIF)

Variable	*Model 1 (Performance)*	*Model 2 (Non-broadcast)*	*Model 3 (Broadcast)*
Weight	2.67		
Appearances	2.81		
Substitutions	1.73		
Goals	3.53	2.76	2.76
Assists	1.56		
Yellow cards	1.40	1.01	1.03
Red cards	1.05		
Shots	3.30	2.74	2.84
Passes	1.73		
Aerials	2.65		
Position	1.45		
Age	1.16		
Height	3.14		
Foot	1.05		
Followers		1.07	2.01
Mentions			2.23

Broadcast Mentions by Clubs	2.18

Table A.5

Transformation of Position dummy variable

Position	*Dummy Variable*
Attacking Midfield	1
Central Midfield	2
Centre-Back	3
Centre-Forward	4
Defensive Midfield	5
Keeper	6
Left Midfield	7
Left Wing	8
Left-Back	9
Right Midfield	10
Right Wing	11
Right-Back	12
Secondary Striker	13

Table A.6

Transformation of Foot dummy variable

Foot	Dummy Variable
Left	0
Right	1

Not defined	2

Table A.7

Twitter club accounts data description summary.

Club	Mentions in media	Number of players	Number of followers	Final standing
Leicester	118167	23	259 793	1
Liverpool	107221	34	9'316 725	8
Arsenal	90107	25	914 365	2
Chelsea	80371	28	1’073 728	10
Manchester United	72354	33	1’0367 156	5
Tottenham	43117	24	464 969	3
Everton	40151	31	933 458	11
Newcastle United	37706	31	4’063 464	18
West Ham	36598	28	1’864 438	7
Manchester City	28220	25	8’101 524	4
Bournemouth	25764	28	1’001 006	16
Southampton	25629	26	6’951 363	6
Sunderland	25344	31	688 928	17
Aston Villa	24004	27	299 259	20
Swansea	23955	27	1’046 020	12
Watford	22350	25	581 512	13
Crystal Palace	17573	30	481 507	15
West Bromwich Albion	16669	28	660 866	14

Stoke City	10522	27	712 443	9
Norwich	5587	28	687 573	19

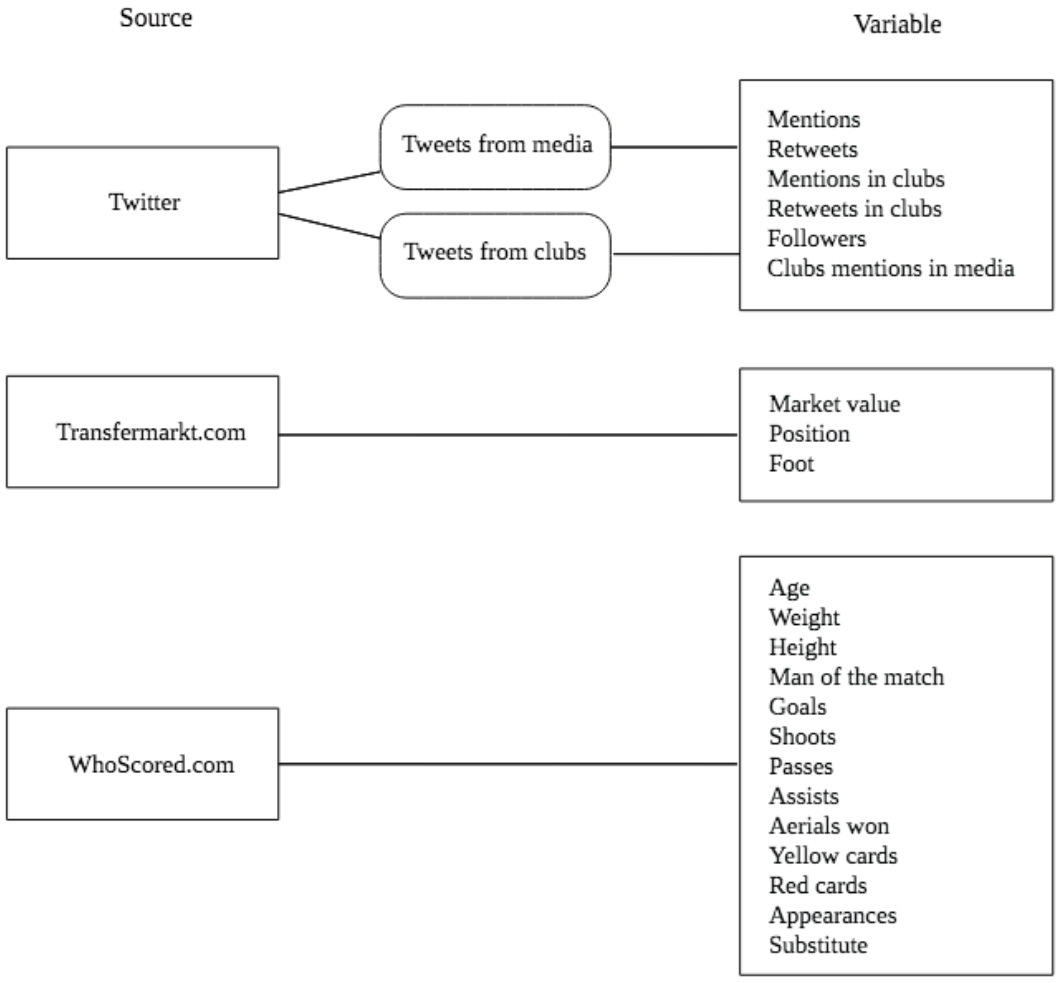

Figure A.1 Sources and corresponding variables.