



INSTITUTO  
UNIVERSITÁRIO  
DE LISBOA

---

Football players performance analysis and formal/informal media:  
Sentiment Analysis and Semantic Similarity

Gustavo Henrique de Sousa Silva

Master Degree in Information Management Systems

Supervisor:

PhD Rui Jorge Henriques Calado Lopes, Associate Professor,  
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,  
Iscte - Instituto Universitário de Lisboa

November, 2021





TECHNOLOGY  
AND ARCHITECTURE

---

Department of Information Science and Technology

Football players performance analysis and formal/informal media:  
Sentiment Analysis and Semantic Similarity

Gustavo Henrique de Sousa Silva

Master Degree in Information Management Systems

Supervisor:

PhD Rui Jorge Henriques Calado Lopes, Associate Professor,  
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,  
Iscte - Instituto Universitário de Lisboa

November, 2021



*S.D.G.*



## Acknowledgment

I could not finish this work without acknowledge those who have a special participation on this work and in my life.

First of all, I would like to thank God, who gave me the ability to work on this dissertation.

My father Luiz, my mother Joana, my sister Juliana and my brother-in-law Davi, without my family I could not achieve this step on my life. You are synonym of hard work and you are my support and the foundation of my life.

Thanks to my grandfather Francisco(in memorian), who is one of my greatest inspirations, – “Faith and courage.” this saying will lead me until the end of my life.

Thanks to all my friends for their support and prayers, I could not reach this step of my carrer without you.

Thanks to my supervisors, Rui Jorge Lopes and Ricardo Ribeiro, for all the support, patience and ability to extract the best work that I could perform.





## Resumo

Nos times de futebol, não diferente com o que já acontece em outros tipos de atividade no moderno mundo dos esportes, tenta-se obter, o quanto possível, informações consistentes dos atletas. Atualmente, é muito importante não só controlar os indicadores fisiológico, nutricional e de saúde, mas também outros aspectos.

A performance de um jogador pode ser medida de um modo objetivo (e.g. Gols marcados, assistências, interceptações). Isso tem sido um método de comparar e ordenar os melhores jogadores por categorias. Após anos de estudo, muitos outros fatores que podem influenciar a performance dos jogadores têm sido descobertos e estudados, considerando não só os fatores objetivos, mas também os fatores subjetivos. Comentários de jogo extraídos de diferentes fontes (e.g., mídia social e mídia formal) também desempenham um importante papel na avaliação de performance subjetiva.

Através da similaridade semântica este estudo busca contribuir com o entendimento de conceitos usados em comentários, especificamente cada palavra-chave associada aos processos do jogo usadas nos comentários publicados nas mídias social e formal.

Esse trabalho também busca analisar o sentimento sobre os times, jogadores e técnicos nas mídias sociais, através da perspectiva dos fãs e da mídia formal sobre a performance. Usando reconhecimento de entidade mencionada e ferramentas de análise de sentimento sobre os comentários e opiniões expressas pelos fãs no Reddit e mídia especializada nos sites de esportes sobre um jogo da UEFA Champions League foi possível distinguir diferentes sentimentos sobre diferentes entidades (jogadores, técnicos e times) e relacionar com aspectos objetivos em um jogo. Além desses resultados também foi possível identificar várias deficiências dessas ferramentas neste contexto. (e.g., ironia e gírias).

Os resultados obtidos mostraram que semanticamente os fãs e a mídia formal comentam sobre o jogo de uma forma mais objetiva, segundo a análise de similaridade feita na estrutura Work domain Analysis (WDA) e através da análise de sentimento foi possível ter impressões sobre os principais tópicos comentados e a relação com alguns eventos do jogo.

Palavras-chave: Futebol, Work Domain Analysis, Match Annotation, Análise de Performance, Análise Semântica, Análise de Sentimento



## Abstract

Football teams, not different from what already occurs in other kinds of activities in the modern sports world, try to obtain, as much as possible, consistent information on athletes. Currently, it is very important to control not only the physiology, nutrition, and health indicators, but several other aspects.

Player's performance can be measured in an objective way (e.g., Goals scored, assists, interceptions), this being seldom a method to compare and rank the best players by categories. Over years of study, many other factors that can influence the players performance have been discovered and studied, considering not only objective factors, but also subjective factors. Match commentary from different sources (e.g., social and formal media) also plays an important role on a more subjective performance assessment.

By using semantic similarity analysis this study aims to contribute to the understanding of the concepts that are used in commentaries, notably key phrases associated to match processes used in entries published in social and formal media.

This work also aims to analyse the sentiment about teams, players, and coaches in social media, thus, it explores the fans' and specialised media perspective on performance. By using named entity recognition and sentiment analysis tools over the set of comments and opinions expressed by fans on Reddit and specialized media on sports sites about a UEFA Champions League match it was possible to distinguish different sentiments on different entities (players, coaches, teams) and relate that with objective aspects of that match. In addition to these results, it was also possible to identify several shortcomings of the usage of these tools in this context (e.g., usage of irony, slang).

The obtained results showed the comments of fans and formal media about a match as presented in the Similarity Analysis of the Work Domain Analysis (WDA) structure. Through the Sentiment Analysis it was possible to describe some impressions about the most commented topics and the relation with some match events.

Keywords: Football, Work Domain Analysis, Match Annotation, Performance Analysis, Semantic Similarity, Sentiment Analysis



## Contents

Acknowledgment	iii
Resumo	v
Abstract	vii
List of Figures	xi
List of Tables	xiii
List of Abbreviations	xv
Chapter 1. Introduction	1
1.1. Motivation and Topic Relevance	2
1.2. Questions and Research Goals	2
1.3. Methodological Approach	3
1.4. Structure of Dissertation	3
Chapter 2. Literature Review	5
2.1. Performance Analysis	5
2.1.1. Match Annotation	6
2.1.2. Work Domain Analysis	7
2.2. Football and Social Media	9
2.3. Tools	10
2.3.1. Similarity Analysis	10
2.3.2. Sentiment Analysis	10
Chapter 3. Research Methodology	13
3.1. Materials	13
3.1.1. Dataset used for Similarity Analysis	14
3.1.2. Dataset used for Sentiment Analysis	15
3.1.3. Reddit	16
3.1.4. Sports Site Live Comments	17
3.1.5. Player Ratings	17
3.2. Methods	18
3.2.1. Similarity Analysis	18
3.2.2. Sentiment Analysis	19
Chapter 4. Results and Analysis	21

4.1. Similarity Analysis	21
4.2. Sentiment Analysis	22
4.2.1. Global	23
4.2.2. Reddit	27
4.2.3. Player Ratings	30
4.2.4. Sports Sites Comments	33
Chapter 5. Conclusions and Future Research	41
5.1. Global Perspective	41
5.2. Semantic Similarity	41
5.3. Sentiment Analysis	42
References	45

## List of Figures

1	Performance Analysis of a player during the match	6
2	Typical analysis of extraction of match annotations in the professional teams	7
3	Sample of Match Annotation - Site Who Scored	8
4	Example of means-end links WDA	9
5	Main goals of teams on Social Media	10
6	Example of representation of semantic relatedness (shorter line represents greater relatedness) [28]	11
7	Reddit users by country [32]	14
8	Reddit Comment Example	16
9	Sports Site Live Comments Example	17
10	Player Ratings Comments Example	17
11	Representation of Similarity Analysis with BERT	19
12	Similarity score between entry and key sentence at different levels ( <i>L4.Functional purposes, L3.Value &amp; priority measures, L2.Purpose-related functions, L1.Object-related processes</i> )	22
13	Global Polarity TextBlob/Stanza	27
14	Global Subjectivity TextBlob	27
15	Players Ratings (WhoScored.com) - UEFA Champions League Final	28
16	Sports Site Live Comments Example	28
17	Top 10 commented Entities and MLE Analysis - Reddit	29
18	Top 10 commented Proper Nouns and MLE Analysis - Reddit	29
19	Entity Polarity TextBlob - Reddit	30
20	Proper Nouns Polarity TextBlob - Reddit	30
21	Entity Subjectivity TextBlob - Reddit	31
22	Proper Nouns Subjectivity TextBlob - Reddit	31
23	Entity Polarity Stanza - Reddit	31
24	Proper Nouns Polarity Stanza - Reddit	32
25	Top 10 commented Entities and MLE Analysis - Player Ratings	32
26	Top 10 commented Proper Nouns and MLE Analysis - Player Ratings	33

27	Entity Polarity TextBlob - Players Ratings	34
28	Proper Nouns Polarity TextBlob - Players Ratings	34
29	Entity Subjectivity TextBlob - Players Ratings	34
30	Proper Nouns Subjectivity TextBlob - Players Ratings	35
31	Entity Polarity Stanza - Players Ratings	35
32	Proper Nouns Polarity Stanza - Players Ratings	35
33	Top 10 commented Entities and MLE Analysis - Sports Sites	36
34	Top 10 commented Proper Nouns and MLE Analysis - Sports Sites	36
35	Entity Polarity TextBlob - Sports Sites Comments	37
36	Proper Nouns Polarity TextBlob - Sports Sites Comments	37
37	Entity Subjectivity TextBlob - Sports Sites Comments	38
38	Proper Nouns Subjectivity TextBlob - Sports Sites Comments	38
39	Entity Polarity Stanza - Sports Sites Comments	38
40	Proper Nouns Polarity Stanza - Sports Sites Comments	39



## List of Tables

1	Sentiment and emotion comparison [29]	11
2	Threads and number of comments - Reddit	15
3	Number of comments - Formal Media	15
4	Analysed threads and number of comments	15
5	Number of comments - Formal Media	16
6	Similarity of the different information sources with the WDA levels (mean and standard deviation)	23
7	Comparison of rank and mean across layers and entity sources - Object Related Processes	24
8	Comparison of rank and mean across layers and entity sources - Purpose-Related Functions	25
9	Comparison of rank and mean across layers and entity sources - Value & Priority Measures	25
10	Comparison of rank and mean across layers and entity sources - Functional Purposes	26



## List of Abbreviations

**MLE:** Maximum likelihood Estimation.

**WDA:** Work Domain Analysis.



## CHAPTER 1

### Introduction

*"Anyone who wants to succeed in any activity must identify and understand the logic behind it, reinterpret it and adapt it to new realities and challenges."* [1]

Football (Association football or soccer) is a worldwide passion and is an industry that always moves the economy with astronomical values, football has been more demanding everyday, both in relation to the structure of the clubs and in relation to the preparation of their squads.

Sports economics articles appear in leading the economic journals and most economists agree that together with the social and cultural importance, professional sports is an area in evidence of both theoretical and empirical research [2].

According to Cotta [3], analysing the opponent and being able to better prepare a match is totally related to the help of a professional who takes care of these issues and leaves everything ready for the coach, technical committee and athletes to receive filtered and punctual information about their team and opponent.

In football, the team with the best quality does not always wins, but the truth is that the winner is always the one who knows how to take advantage of opportunities, regardless of their physical, technical, and economic reality. What is happening in the current scenario is that the teams try to align their reality with the most likely opportunities for success within a football match, and for this, the statistical information and the analysis of the games' performance are increasingly in evidence and those who can extract more data in an efficient and effective way have an advantage in the search for positive results within the games.

Team and athlete performance analysis has been an object of study and usage by practitioners (e.g., coaches) for several years. Methodologies, metrics, and studies have been designed to improve the performance of football players and provide a better performance analysis, typically in an objective way using notational analysis to account for several athletes actions (e.g., goals, assists, shoots).

The combination of human factors and football complexity makes performance analysis an extremely challenging task. Advances in these studies provide an increasing number of factors that are considered to influence players' and teams' performance. On the other hand, the perceived performance, e.g., by fans or even specialised media, generally does not follow these procedures and metrics and is not expressed via objective metrics.

This dissertation leverage on these facts by assessing the subjective performance of the football players, based on the context of the match, using the comments of users in social media and the specialised media opinion on sports sites. Notably, we perform

named entity recognition and sentiment analysis on these comments in order to understand if the subjective perception by fans and formal media has relation with the objective performance on the pitch.

### **1.1. Motivation and Topic Relevance**

The motivation for choosing the theme is due to the interest in football and the integration of technology in the sports scenario.

This study aims to explore the gap that may exist between objective performance approaches and their metrics and the subjective performance assessment expressed by fans in social media. Being the most popular sport in the world, football and football players have attracted much attention and followers in the social media [4]. Consequently, athletes are armed with highly effective online communications tools that enable them to garner the influence like the great corporations have [5]. In fact, the importance of reputation is chief and has a direct impact on players' and teams' performance.

Seeking to align the dissertation with the author's research, it was decided to investigate a subject that has been studied in various spheres of society. Sentiment analysis has proven essential to get insights on a certain aspect based, not on formal surveys, but on what people say in everyday environments such as social networks.

Social networks indicate much about the users' preferences, it is possible to recognise people's profiles, from their material choices such as: cell phone brands, places they usually visit in their daily routines, main brands of clothes they like to wear; to personal choices such as: political orientation and opinion on sensitive subjects. Moreover, when it comes to football, social networks are very active. Both with messages supporting the team and complaining in case of adverse results.

Possibly the greatest value of football clubs is in their fans: the more satisfied the fans, the more they give prestige to the club. Social media can have important information to monitor the sentiment of the fans about a team and the respective players.

### **1.2. Questions and Research Goals**

The general research question addressed in this work is the following:

- What is the relationship between the formal media commentaries and social network users' feelings towards players during matches?

Nowadays it is possible to measure the players' indicators, such as shots, tackles, dribbling, among other statistics that are fundamental to get evidence of performance, their strengths, and the fundamentals that need to be improved. Football carries a factor that is what makes it such a popular sport, there are players more skilled than others, teams more organised than others, but during the ninety minutes, anything can happen: the weaker can beat the stronger, the less skilled can be the differential factor for a victory.

The emotion that football generates in people's lives is often expressed on social networks. Through an analysis about a particular club or player, it is possible to get some clues about the behaviour that the fan has in relation to the sport.

The present work seeks to cover the following objectives:

- Verify the relationship between formal media perception of players during matches and the analysis of fan sentiment on social networks during those same matches;
- Explore the perception concerning the usage of Work Domain Analysis (WDA) structure of Football

### **1.3. Methodological Approach**

The present work will use the following methodology:

- Contextualisation of the state of the art and literature review;
- Presentation of the materials used for this study;
- Extraction and mining of the data to be studied;
- Development of a method to relate the object of study – on the one side the statistics extracted from the football players; and, on the other side the publications made by the users of social networks and specialised media;
- Analysis of the results obtained by comparing objective data and the social network users data.

### **1.4. Structure of Dissertation**

This study is organised into five chapters that aim to contextualise the theme and present the methodology and results. The first chapter seeks to contextualise the theme and highlight the importance of this field of research, in addition to addressing what specifically is going to be studied and in what way. The second chapter deals with the theoretical framework, called Literature Review, and is separated into three sections, important themes that are totally linked to the study carried out.

The third chapter presents the material and the sources analysed and the method of how this study was conducted.

The fourth chapter shows the results and the respective analysis and a contextualisation between the objective performance and the subjective performance.

The fifth chapter presents the conclusions and perceptions of this study and the possibilities of future research to improve the results and extend the coverage of this project.





## Literature Review

### 2.1. Performance Analysis

There are many factors that can influence a match result. Over the years, researchers have been trying to analyse the complexity of these factors. Many aspects of human behaviour can be analysed, consequently, it is important to determine what will be analysed and the reason for this. It is important to consider the saying: “not everything that counts can be counted, and not everything that can be counted counts” [6]. The sentence above defines what happen in a football match. Players can have a bad performance considering some stats (e.g., goals, assists, shots, interceptions), but still make a good match if we consider the match context, for example, a player who was positioned to avoid counter-attacks or marked a specific opponent player individually and played this role positively, both are hard aspects to measure, but that we cannot ignore when analysing the performance.

**Physiological Factors:** Studies suggest that fatigue has a major influence on the frequency of players’ participation with the ball and with the decline in intensity over the course of the game [7], [8]. In addition, muscle fatigue puts players at risk of injury not only from the effort in each match, but from the accumulated effort over the course of the season [9]. Another physiological factor that influences performance is injuries. A survey conducted with some teams in Europe recorded an average of two injuries per season per player in the teams studied [10]. In addition to the factors previously mentioned, the nutrition is also an important factor to be taken into account, what a player ingests in the moments before the game can reduce the impact of fatigue, allowing their performance to improve, as well as what is consumed after the games or during training can help the recovery of the player.

**Psychological Factors:** According to a study by Gouttebarga and Kerkhoffs [11], football players who have suffered one or more serious injuries (which require them to stay a long time recovering) are four times more likely to suffer from psychological problems than other athletes. In addition, the psychological disorders studied have a negative influence on the player’s performance.

**External Factors:** Weather conditions [12], ball pressure [13], type of pitch [14], pitch size and time of match [15], altitude [16] are all factors that are not directly related to the player, but that have been proven to determine performance.

**Strategic Factors:** Coach and staff can manage different functions for a player to perform in a match, it depends on the context. A data analysis from Spanish top division games, suggested that “Top3” teams earned more assists, shots on goal, touches on the ball, passes, threaders, dribbles, and successful long passes, but fewer tackles than “Bottom3” teams. Another aspect of the study is that the defenders and fullbacks of the better teams participated in more offensive actions than those of the worse teams, but conversely, participated in fewer defensive actions. The statistical results of a player are linked to the role he/she plays in each game, the quality of the opponent and the strategic context of the match [17], [18].

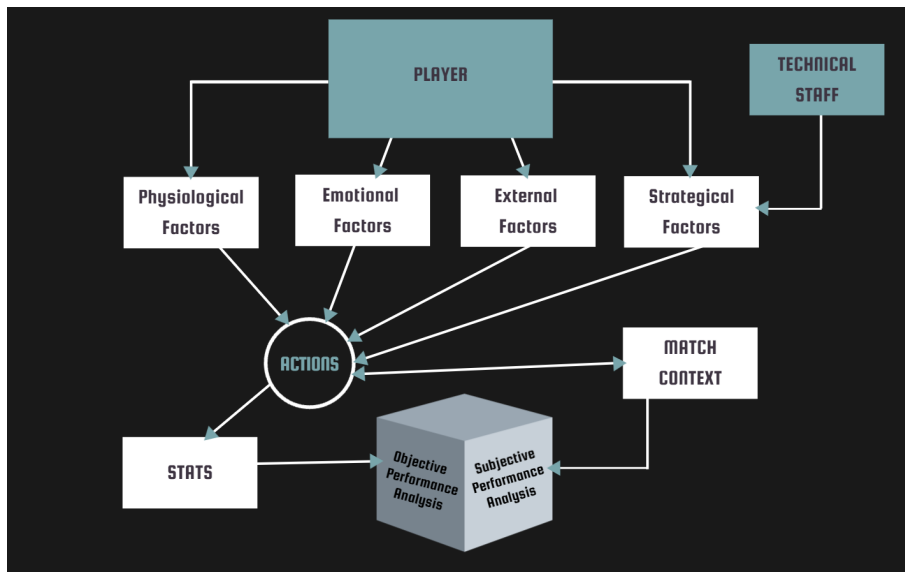


FIGURE 1. Performance Analysis of a player during the match

As illustrated in Figure 1, the player’s actions are conditioned by the factors indicated above, notably the match context is a main factor in sports performance. In fact, there are many human and non-human components operating dynamically and constantly changing the match environment. The complexity of humans combined with that of football makes the difficulty of analysing player performance high. As time goes on, more and more factors are found that can influence a player’s performance [19].

There are tools(e.g., annotations) that can obtain the objective performance of the match and generate stats that turn possible to compare players through an objective analysis, this dissertation is focused in the other perspective, subjective analysis, based on the perception of the fans and formal media comments.

### 2.1.1. Match Annotation

Annotations in football are an important tool to obtain intelligence in a match, even in a general performance of the team or an individual classification of a player. With the sports evolution on the last decades the need of more researches about the complexity of the evaluation of a player performance was found [20]. In football scenario, even it being

a sport very complex, it is possible to analyse the participation of a player in a match through predefined stats (e.g., shots, interceptions, assists, goals).

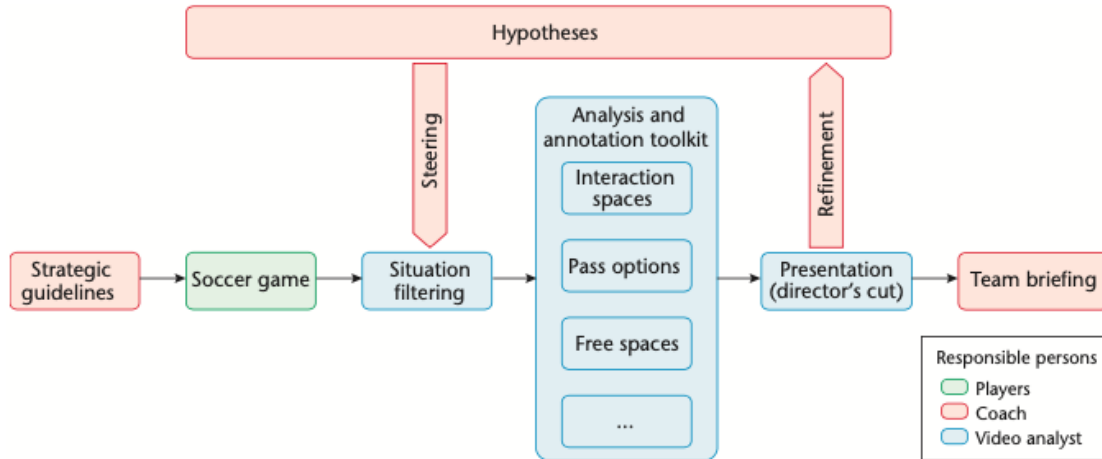


FIGURE 2. Typical analysis of extraction of match annotations in the professional teams [21].

The match annotations are alternatives to analyse a player performance in an objective way, and to compare a player with other player based in a common stat. According to Figure 2, the player participation in a match can be converted in stats by analysis tools and used by coaches and clubs to assist in decision making.

This form of analysis is used not only by football teams nowadays, but also by television channels, sports journalists, betting websites, among others. In the present work we will use the match annotations extracted from the website “whoscored.com”.

As we can see in Figure 3, the site presents statistics from a football match and uses them to inform the performance of individual players and the team overall, so that each foundation of the game can be fit and ranked. Through this information it is possible to know who had the best and the worse performance and obtain impressions about the match events.

### 2.1.2. Work Domain Analysis

Work Domain Analysis (WDA) is a system analysis method that aims to, in a structured mode, associate actors, their fundamental functions and resources used by themselves in a context, based in the functional environment that establishes the purposes to be achieved.

In the football scenario, the whole squad has several common functions (e.g., positioning, connect passes, etc.), but each position has specific roles on a football match (e.g., a striker is the responsible for scoring the goals, the central back is the responsible to intercept the opponents and avoid opponents effective attacks). Based on a preview study, Berber et al. classified hierarchically a conceptual method to link the functions and purposes of players in a football match [22].



FIGURE 3. Sample of Match Annotation - Site Who Scored. Available in: <https://www.whoscored.com/Matches/1544080/Live/Europe-Champions-League-2020-2021-Manchester-City-Chelsea>

The structure is designed from specific components to general components, each level is linked with the adjacent level based on the relation of the purpose and functions of the player position in a match.

**Functional Purpose:** The main functions of a player in a match (Prevent goals scored, Score goals, Relieve pressure, Create chances). Example: a striker has as main function to score goals.

**Values and Priority Measures:** Criteria used to analyse the progress of a player to achieve the functional purposes (Positioning, Goal Conceded, Saves made, Goals scored). Example: the quantity of goals a striker scored in a match.

**Purpose-related Functions:** Functions that need to be done to achieve the functional purposes (Defend, Attack, Leadership, Adaptability, Communication). Example: a striker has to establish communication with the teammates to find the best way to score goals.

**Object-related Processes:** The process used by players to achieve a purpose-related function (Dive, Shooting, Break Lines, Free Kicks, Vision). Example: a striker has to pass, tackle, and kick to perform the purpose-related functions.

**Physical Objects:** Objects and external factors used by players to practice the football (Boots, Gloves, Playing Kit). We not use this level on this study, because it is not totally related with the players action and consequently with the performance analysis of a player.

In Figure 4 we can observe an example of WDA means-end links of a goalkeeper and the key phrases separated for each level. Through this example, it is easy to understand how the five levels of WDA detail each layer of a work and together build a global concept of the work.

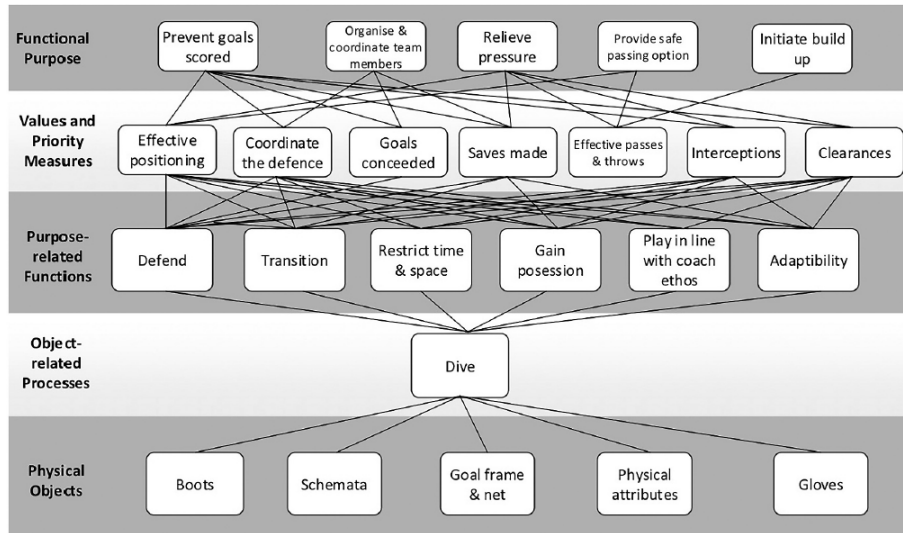


FIGURE 4. Example of means-end links WDA - Goalkeeper extracted from [22]

## 2.2. Football and Social Media

Football is the most popular sport in the world, football players attract a lot of attention and followers on social media [4]. There are studies about brand management [23], relationship between team and fans [24], strategy to attract new fans [25] that deal with this relation of football with social networks and how nowadays it is important to have a good management of these networks to obtain popularity and gain more and more visibility.

Football clubs have two main objectives on social media: to attract people to their official publications and to communicate directly with fans (Figure 5). Having a closer and more direct communication with fans is of great value to the club, there is a feeling that social media platforms have the ability to “break down walls” between clubs and fans.

Not just a great communication tool, social media is also a great platform for data extraction. The evolution of the media allows a much greater proximity to the public not only to communicate, but also to understand the feelings of the fans. It is possible to obtain in real time, if the fans are satisfied/unsatisfied with some player, if they have some dissatisfaction with the club, if there is an interest/disinterest in some player hiring. The

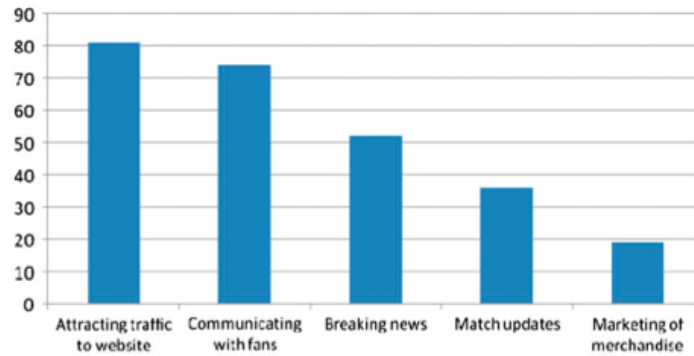


FIGURE 5. Main goals of teams on Social Media [24]

current challenge is to analyse the large volume of data that can be extracted from social networks, preparing it in a well written and structured way, since the data is usually presented with typos, abbreviations, slang, among others [26].

### 2.3. Tools

The following tools were used in the methods to extract data and help to understand examples of how Subjective Analysis works.

#### 2.3.1. Similarity Analysis

Semantic similarity is defined as the measure of semantic equivalence between two blocks of text [27]. Over the time, techniques were developed to refine text comparison, taking into account the semantic relatedness or semantic distance of the words. Semantic distance is the shortest path to link a source word to a target word and the relatedness is an aggregate of all the paths from source word to target word in a semantic space. Two sentences may be close in distance and still not be closely related because there is no aggregate of paths to contextualise the relation of both [28]. In Figure 6, the word “Red” has nodes very near (e.g., “Fire”, “Orange”) that indicate a great relatedness and a short semantic distance between these words and has other distant nodes (e.g., “Sunrises”, “Sunsets) what indicates great relatedness but high semantic distance, otherwise there are nodes that do not have a direct relation with the word “Red” (e.g., “Street”, “Clouds”) what indicates a low relatedness and high semantic distance. The semantic similarity is important for the exploratory work to analyse what is the most commented topics in the formal and informal media related with the WDA structure.

#### 2.3.2. Sentiment Analysis

Anger, admiration, displeasure, sadness, are some of the many emotions that a human being can express. Feelings and emotions are mentioned very similarly in our everyday life, although they are treated differently by psychology. Feeling is the transmission of thoughts derived from an individual’s emotions, what means that a person’s feeling is the way he expresses his emotional state. Represented in Table 1 are the main differences

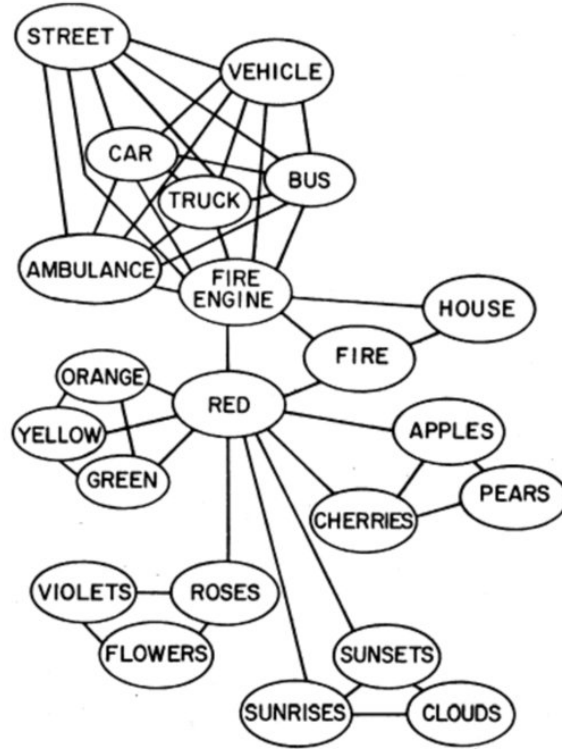


FIGURE 6. Example of representation of semantic relatedness (shorter line represents greater relatedness) [28]

between emotion and sentiment, and how one is more closely linked to the psychological state and the other is much more closely linked to an individual’s behaviour [29].

TABLE 1. Sentiment and emotion comparison [29]

	Definition	Connection	Dimension	Nature
Emotion	Complex psychological states	-	Psychological dimension	Raw and natural
Sentiment	Mental attitudes or thoughts	Expression of emotions	Social dimension	Highly organised

According to Liu [30], sentiment analysis is the field of study that analyses people’s opinions, feelings, evaluations, attitudes and emotions towards entities such as products, services, organisations, individuals, problems, events, topics and their attributes. The greater the volume of data to analyse the sentiment of the audience studied, the greater the possibility of obtaining accuracy about the sentiment related to a topic. Understanding how people feel about some issue makes us more assertive in dealing with problems and gives us a basis for analysing a given issue. There are two main techniques used to do sentiment analysis in an automated way, so they do not depend on human analysis to make a judgement about some written sentence: In the lexicon-based technique, a dictionary of words is used, where each one is associated with a value, and in the end the sum of the values of all the words (usually adjectives) is obtained, thus calculating their semantic orientation (polarity and strength of words, phrases or texts). The second technique is machine learning, where basically it is a supervised classification task on a text, the polarity of a text is defined through classifiers obtained through machine learning [31].

The growth of sentiment analysis studies coincides with the prominent growth of social networks. The growth of these networks leads to a large volume of data containing the opinions of various users on numerous topics, truly functioning as a large data sample[26]. Sentiment analysis was important to understand the perception of the authors in formal and informal media through the calculation of polarity of the comments, and if the content was more objective or subjective. Through this indicators it is possible to analyse the perception of fans and specialised media about the players and teams.



## Research Methodology

The dissertation was separated in two exploratory works, the first approach is a semantic analysis where we verified the similarity between the datasets composed of informal and formal media entries and WDA key phrases. The second approach is focused on sentiment analysis. We studied the polarity and the subjectivity of each entry and relate with some match events and some objective aspects of the match.

### 3.1. Materials

Three different sources were used in this project one of Informal Media(e.g., Reddit) and two of Formal Media(e.g., Live Match and Player Ratings), each source is a different perception of the same match, based on the context of each platform. One of the chosen sources was the social media Reddit. Reddit comments are presented in an informal language, where essentially the author has an open space to express anything he/she wants about a particular topic (e.g., football match). In Formal Media, there are two sources: Live Match commentary and Player Ratings. The first one, Live Match, is the update, in real time, of the events in a match and the comment about the events as they happen. The Player Ratings comments are analyses about the general participation of a player in a match and the respective rating of the performance.

In the present study, Reddit was used for obtaining the social media content data. This platform aims to connect users by grouping them in communities through the creation of rooms about topics, where users can comment and react to other's comments. In 2020, Reddit had over 52 million daily active users, nearly 303 million posts and two billion comments. With all this comments and reactions, Reddit now is more than a simple social media, it is a massive data repository, where we can retrieve and analyse comments on about almost any topic in the world. Users around all the world are using Reddit, this number is increasing year over year. Portugal is ranked in the top 10 of the users by country, as shown in Figure 7. Reddit is organised around the following concepts:

**Users:** who interact in Reddit. A user can comment and react in threads, follow other users and join communities.

**Communities:** Group of *Users* with common interests about a topic.

**Threads:** A room where users can interact about a given topic, for example, in a Football Match Thread, the principal objective is talk about football and related subjects.

**Comments:** A space designed for users to interact with other users or just to express opinion. *Comments* must respect the platform policies, but, the user is free to express his opinion using any language, including slang, emojis and hashtags.

Reddit offers an Application Programming Interface (API). By using it, we can retrieve the data through authenticated requests. Then, this data can be filtered and organised. In the end, it can be used to study topics or the sentiment associated to this kind of contents. Also, through Reddit API it is possible to make filtered searches about *Users*, *Communities* and *Threads*. Using the Reddit API, a dataset containing fans' comments and opinions on that match was created.

To obtain formal media contents, it was used web scraping, that consists in collecting data from web pages to obtain data, from different sources, in this case, sports sites.

As case subject, we used the match between Manchester City F.C. and Chelsea F.C. on the final of the UEFA Champions League 2021, that took place on May 29, 2021. In this match, Chelsea F.C. beat Manchester City F.C. by 1-0, winning the tournament.

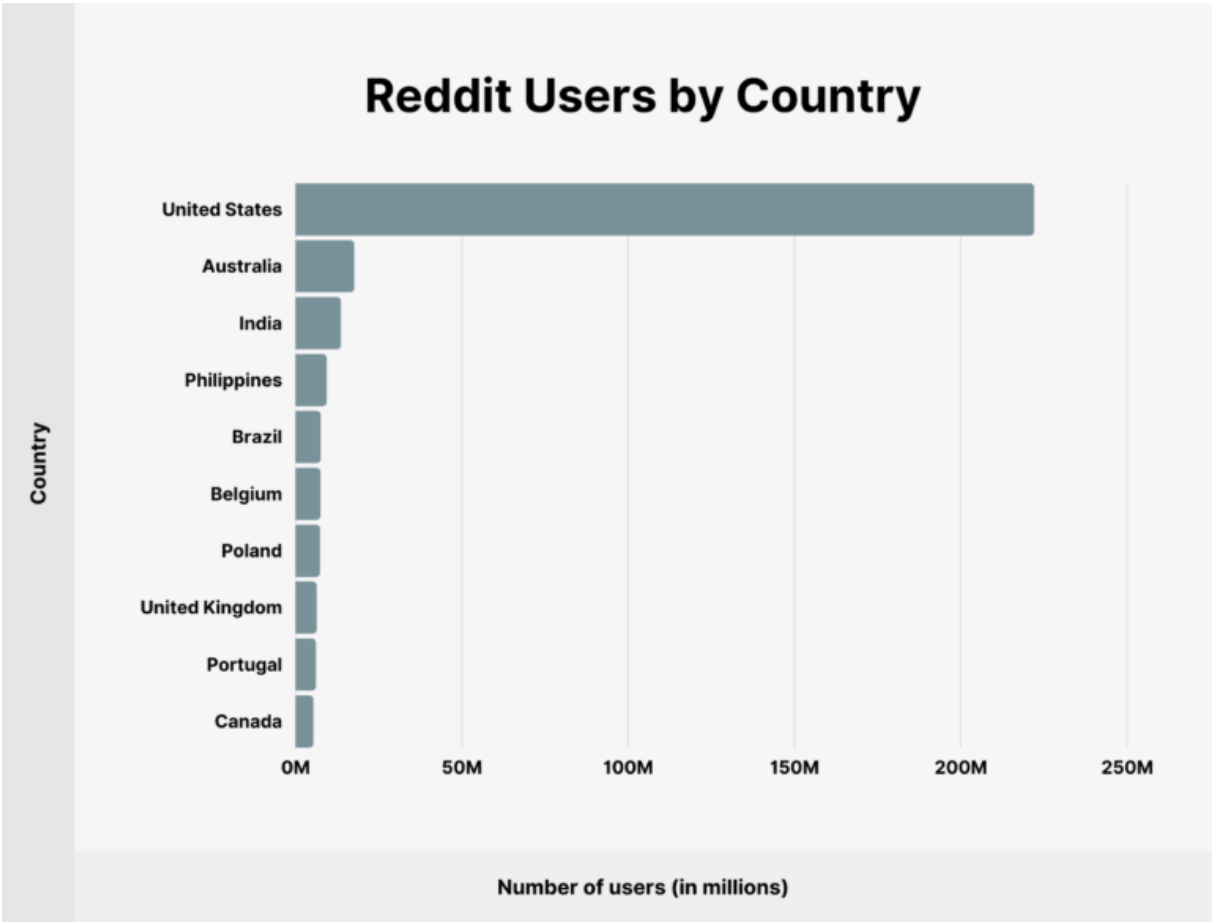


FIGURE 7. Reddit users by country [32]

### 3.1.1. Dataset used for Similarity Analysis

As described in Table 2, we have a total of 1164 comments extracted from Reddit. To a better perception we separated the comments in sentences, and the result obtained was the total of 2017 sentences. This dataset followed a random criteria of selection based on the Reddit API search.

TABLE 2. Threads and number of comments - Reddit

Thread	#Comments	#Authors	Word Avg.
[Match Thread] Manchester City vs Chelsea (Champions League final)	470	386	16.80
[Post-Match Thread] Manchester City 0-1 Chelsea (Champions League final)	121	116	26.75
[Match Thread] Manchester City vs Chelsea (UEFA Champions League Final)	488	255	18.78
[Pre-Match Thread] Manchester City vs Chelsea (Champions League final)	85	79	22.47
Total Comments	1164		

TABLE 3. Number of comments - Formal Media

Thread	#Comments	#Sites	Word Avg.
Sports Site Live Events	257	5	50.6
Player Ratings	331	19	34

As described in Table 3, we have a total of 257 comments of Sports Sites Live Events and 331 comments of Player Ratings, to a better perception we separated the comments in sentences, and the result obtained was the total of 711 sentences and 802 sentences respectively. On the Semantic Analysis we separated the comments in sentences, because a single comment can be similar with a range of key phrases, splitting these comments means that we can be more specific with key phrases identification in the text.

### 3.1.2. Dataset used for Sentiment Analysis

TABLE 4. Analysed threads and number of comments

Thread	#Comments	#Authors	Word Avg.
[Match Thread] Man City vs Chelsea   UCL Final	478	305	15.22
[Post-Match Thread] Man City 0-1 Chelsea   UCL Final	317	273	20.99
[Pre-Match Thread] Manchester City vs Chelsea   UCL Final	258	215	37.20
[Pre-Match Thread] Manchester City vs Chelsea   UCL Final	252	184	26.10
[Match Thread] Manchester City vs Chelsea (UEFA Champions League Final)	486	248	18.71
Total Comments	1791		

As described in Table 4, we have a total of 1791 comments, we can see an average number of nearly 358 comments per thread. This dataset followed a random criteria of selection based on the Reddit API search.

TABLE 5. Number of comments - Formal Media

Thread	#Comments	#Sites	Word Avg.
Sports Site Live Events	257	5	50.6
Player Ratings	331	19	34

### 3.1.3. Reddit

The dataset is the result of the following search queries: “match thread city chelsea ucl” and “champions league match thread manchester city chelsea”. The final output structured in accordance to the following pattern:

- A comment made by an user in a thread is analysed;
- Entities are detected by Spacy and highlighted in the comment – there are many types of entities, like Organisation, Location, Person, and Date. In the Figure 8 the given entities are highlighted in green;
- Proper nouns are detected and highlighted in the comment. In the Figure 8 the given entities are highlighted in black;
- Polarity and subjectivity are computed using Stanza and TextBlob.

I am never, ever going to be okay with **Sergio** teary after that game.  
**God** damn that hurts.

**Chelsea** earned the win after **Pep** made the same key mistake for at least the **third** time in big **CL** games.

\*\***Ilkay**. **Gundogan**. Is. Not. **A.** **Defensive**. **Midfielder**\*\*

Especially not against top competition. **Gundogan** way out of position opened the passing lane that led to the goal. **Gundogan** out of position left our defense exposed for far, far too long. This is at least the **third** time that **Pep** has changed his lineup to trust **Gundogan** as a defensive mid on a key **CL** night and it's cost us the game.

And how many times have we all seen a glaring tactical error not **20 minutes** into a big game, and it inevitably takes **Pep** until **65 or 70 minutes** to make the substitution that fixes it?

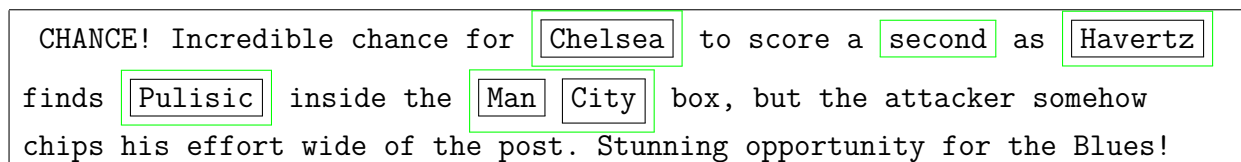
I just don't have words for how frustrating this. It's not just making a mistake in our **first** **CL** final. It's that it's the exact same **two** mistakes **Pep** has made before.

FIGURE 8. Reddit Comment Example

### 3.1.4. Sports Site Live Comments

The dataset is the result of the text wrapping of the related Sports Sites. The final output structured in accordance to the following pattern:

- A real time event published is analysed;
- Entities are detected by Spacy and highlighted in the comment – there are many types of entities, like Organisation, Location, Person, and Date. In the Figure 9 the given entities are highlighted in green;
- Proper nouns are detected and highlighted in the comment. In the Figure 9 the given entities are highlighted in black;
- Polarity and subjectivity are computed using Stanza and TextBlob.



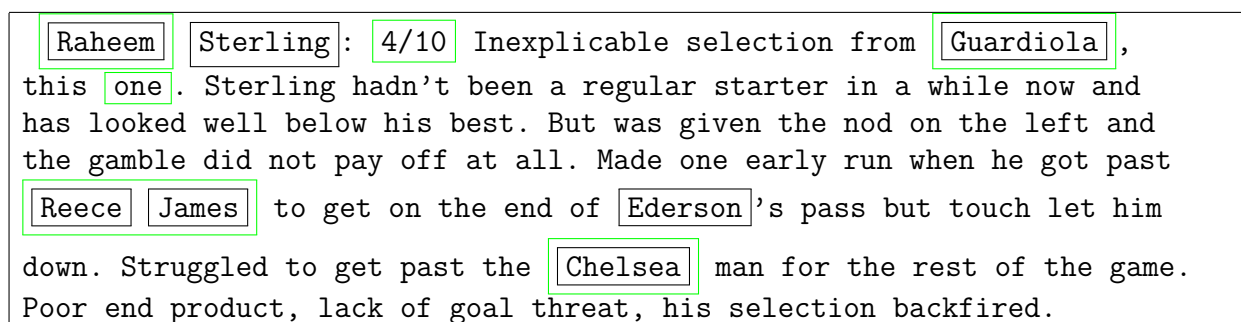
CHANCE! Incredible chance for Chelsea to score a second as Havertz finds Pulisic inside the Man City box, but the attacker somehow chips his effort wide of the post. Stunning opportunity for the Blues!

FIGURE 9. Sports Site Live Comments Example

### 3.1.5. Player Ratings

The dataset is the result of the text wrapping of the related Sports Sites. The final output structured in accordance to the following pattern:

- Each player ratings comment is analysed;
- Entities are detected by Spacy and highlighted in the comment – there are many types of entities, like Organisation, Location, Person, and Date. In the Figure 10 the given entities are highlighted in green;
- Proper nouns are detected and highlighted in the comment. In the Figure 10 the given entities are highlighted in black;
- Polarity and subjectivity are computed using Stanza and TextBlob.



Raheem Sterling: 4/10 Inexplicable selection from Guardiola, this one. Sterling hadn't been a regular starter in a while now and has looked well below his best. But was given the nod on the left and the gamble did not pay off at all. Made one early run when he got past Reece James to get on the end of Ederson's pass but touch let him down. Struggled to get past the Chelsea man for the rest of the game. Poor end product, lack of goal threat, his selection backfired.

FIGURE 10. Player Ratings Comments Example

As described in Table 5, we have a total of 257 comments of Sports Sites Live Events and 331 comments of Player Ratings. On the Sentiment Analysis, we decided to analyse the entire comment, because the comments are a set of opinions of the authors and

the context is a construction of the sentences, what means that the sentiment is not independent from one sentence to another.

## 3.2. Methods

With the material previously presented we used Similarity Analysis and Sentiment Analysis to compare objective and subjective data. In Similarity Analysis we compare the WDA key phrases with the media entries and in the Sentiment Analysis we measure the polarity and the subjectivity of each entry.

### 3.2.1. Similarity Analysis

To understand the relation between the perceived performance by fans and specialised media, we compute the semantic similarity between the Reddit’s posts, live comments, and players’ assessments and the levels of WDA. We experimented different approaches and the best results were achieved by using BERT [33] to generate computational representations of the textual data from fans and formal media and the key phrases corresponding to several levels of WDA. We use the cosine (Eq. 3.1) to compute the similarity between these vectorial representations.

$$\text{sim}_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \quad (3.1)$$

BERT stands for Bidirectional Encoder Representations from Transformers, which hints about its nature. BERT is a language representational model which uses context, left and right, to generate representations for raw text. This model is based on the concept of transformer, which is a neural network architecture that follows the encoder-decoder structure *using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder* [34]. In this work, we used DistilBERT [35], a more efficient version of BERT, that achieves comparable results. As implementation, we used the Python Sent2Vec<sup>1</sup> package.

In Figure 11 there is a representation of the comparison between a Reddit entry and two WDA key phrases

The method to obtain the similarity analysis of the data sources follows three steps:

- (1) **Comments Process** We start by processing each comment and subdividing in sentences. The same comment can have more than one sentence, we separated the comments in sentences, because each sentence can be similar to different key phrases of WDA.
- (2) **Similarity Analysis** After the processing of the comments, we used BERT to compare the sentences with the WDA key phrases, the tool analyses the similarity of the comment in the range of 0 to 1, where 0 indicates less similarity and 1 indicates more similarity

---

<sup>1</sup><https://github.com/pdrm83/sent2vec>

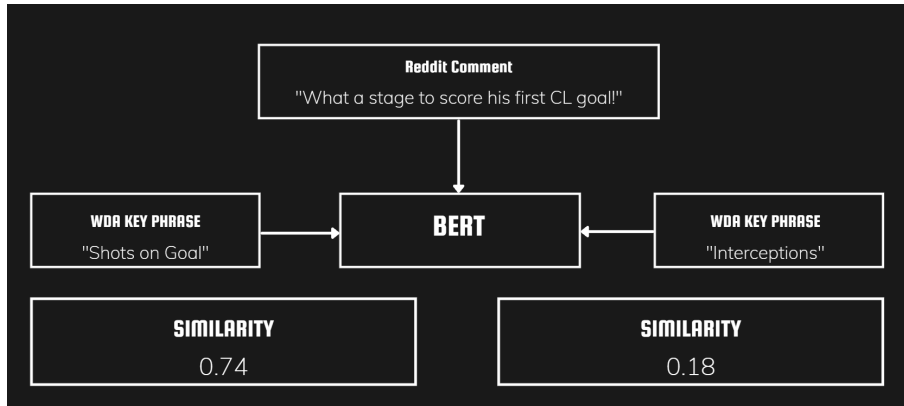


FIGURE 11. Representation of Semantic Analysis with BERT

- (3) **Output Generation** Through this method it was possible to create 12 matrices (corresponding to the three datasets and four levels) with the combinations between key phrases and datasets, we represented each matrix in a heatmap. The color of heatmap corresponds to the values obtained with BERT, where blue is closest to 0 and red is closest to 1, the variation is the result of the method application between each comment entry with each WDA key phrase.

### 3.2.2. Sentiment Analysis

The method to obtain the sentiment analysis of the data sources follows three steps:

- (1) **Linguistic Analysis** We start by analysing each comment. In this step, named entities and proper nouns are recognised to understand the main subjects of the match. When there is repetition of an entity or proper noun a cluster is created, to group common comments.
- (2) **Sentiment Analysis** After the grouping of the entities, the tool analyse the polarity of the comment in the range of -1 to 1, where -1 is extreme negative and 1 is extreme positive, the subjectivity also is analysed in the range of 0 to 1, where 1 is a totally subjective and 0 is totally objective.
- (3) **Output Generation** The polarity/subjectivity of each comment is expressed using a decimal value (e.g., polarity = 0.41) in the previously mentioned range.

On this project, we used the following Python libraries to analyse the comments of Football fans: TextBlob<sup>2</sup>, Spacy<sup>3</sup>, and Stanza<sup>4</sup>. Spacy function is to get the named entities and proper nouns and to separate by type. TextBlob and Stanza libraries are used to analyse the polarity and subjectivity of an expression. With the named entity types we can observe, for example, what is the topic most commented in a thread, what is the polarity about a given football player and the objectivity of the topics.

<sup>2</sup><https://textblob.readthedocs.io/>

<sup>3</sup><https://spacy.io>

<sup>4</sup><https://stanfordnlp.github.io/stanza/>





## Results and Analysis

### 4.1. Similarity Analysis

The method described in Section 3.2 was applied to the collected datasets described in Section 3.1. Specifically, we compute the semantic similarity,  $sim_{ij}^{mn}$  between each entry (i.e., sentence),  $s_i^m$  and key phrase,  $k_j^n$  defined via WDA (here  $s_i^m$  corresponds to the  $i^{th}$  sentence of dataset  $S^{m=1,\dots,3}$ , and  $k_j^n$  corresponds to the  $j^{th}$  WDA key phrase at level  $L^{n=1,\dots,4}$ ). This results in 12 matrices (corresponding to the three datasets and four levels), with values between 0 and 1, which are presented in Figure 12, with each row and column corresponding to entries (sentence) and WDA key phrases respectively. These results show a great dispersion of the similarity score across all domains: i.e., between sources, levels, and between entries from the same source at the same WDA level.

In order to assess how the similarity varied between levels and datasets we computed the similarity mean and standard variation for all the 12 level/dataset combinations. According to Table 6, the general content of all the data sources is more similar with the key phrases from the WDA level *L3. Value & Priority Measures Level*, which means that both, informal and formal media, tend to describe matches using an objective perspective more based on players stats and less based in their participation in more abstract processes (described in level L4). In contrast to this, the WDA level that has less similarity with the content of each information source is *L1. Object-related processes*, which means that in a general context, the comments are not about the secondary (i.e. “means-to-an-end”) functions of a player in a match but about the objective performance and the principal functions (e.g., a striker has to score goals). On the other dimension, at all four levels, Reddit entries present the higher similarity values while Ratings present the smaller values. This is contrary to what was expected, that is, that formal media live commentary and player ratings would be more semantically similar to WDA key phrases than fan’s comments on social media.

We also investigated if the key phrases at each level would or not maintain their similarity rank across the different data sources. According to Tables 7, 8, 9 and 10 the ranking of most similar key phrases is very similar in the three data sources, i.e., informal and formal media comments typically tend to comment the match based on the same key phrases, there is not a great dispersion comparing all the data sources in all the levels.

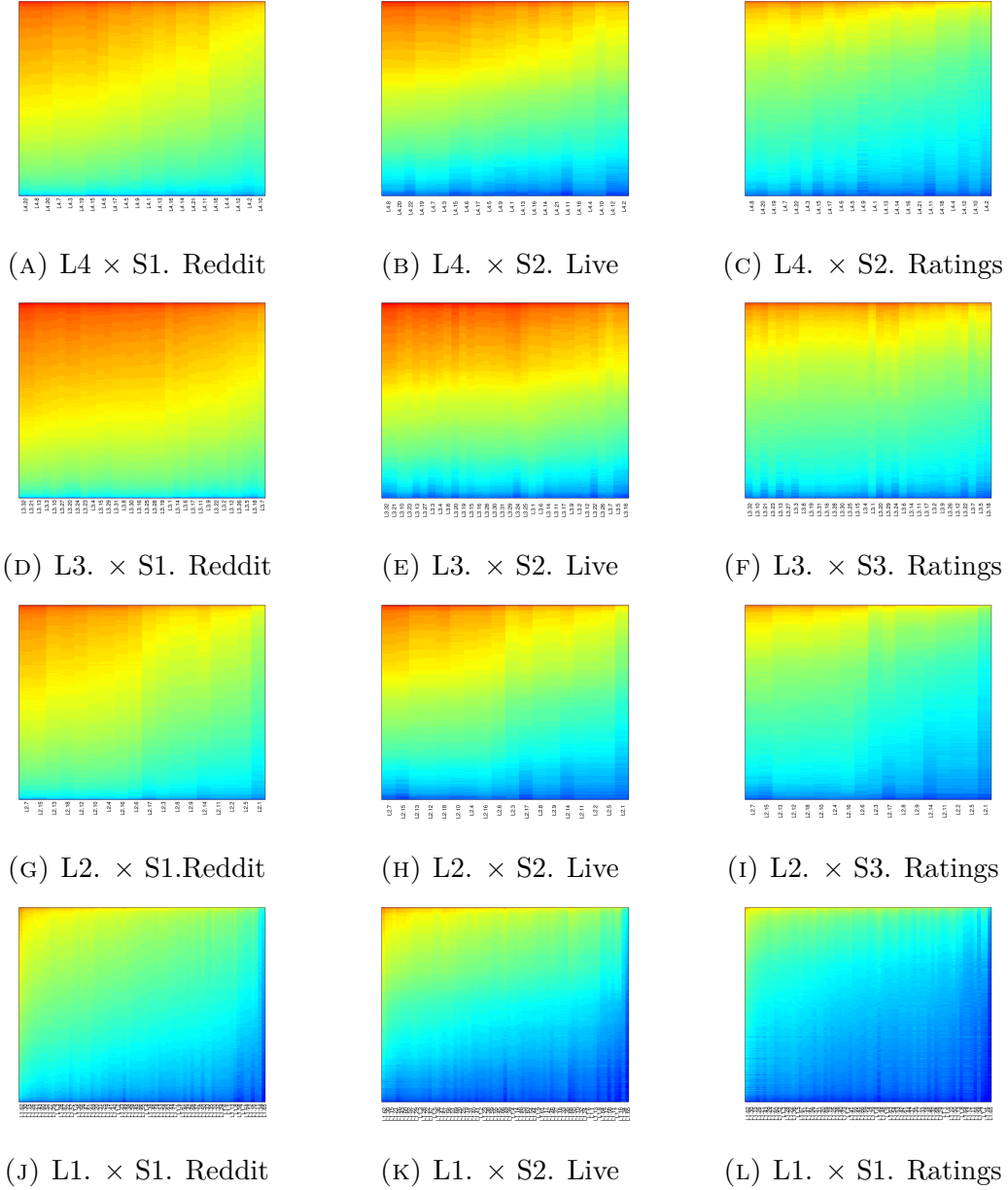


FIGURE 12. Similarity score between entry and key sentence at different levels  
*(L4.Functional purposes, L3.Value & priority measures, L2.Purpose-related functions, L1.Object-related processes)*

## 4.2. Sentiment Analysis

To understand the results of this exploratory work, it is important take into account some concepts, used to evaluate the sentiment and to detail how the information was analysed:

- Polarity: Is the measure of positivity or negativity of a comment, we used two libraries to calculate the polarity (TextBlob and Stanza) and the result starts from -1 (extremely negative) to 1 (extremely positive).

TABLE 6. Similarity of the different information sources with the WDA levels (mean and standard deviation)

	S1. Reddit
L4.Functional purposes	0.437±0.078
L3.Value & priority measures	0.468±0.078
L2.Purpose-related functions	0.411±0.078
L1.Object-related processes	0.326±0.075
	S2. Live Commentary
L4.Functional purposes	0.399±0.104
L3.Value & priority measures	0.427±0.107
L2.Purpose-related functions	0.378±0.099
L1.Object-related processes	0.300±0.089
	S3. Ratings
L4.Functional purposes	0.351±0.081
L3.Value & priority measures	0.378±0.083
L2.Purpose-related functions	0.330±0.077
L1.Object-related processes	0.253±0.067

- **Subjectivity:** Is the measure of subjectivity of a entry, if the author based the comment in factual information or in a personal opinion. We used the library TextBlob to calculate the subjectivity and the results starts from 0 (extremely objective) to 1 (extremely subjective).
- **Entity:** With the Named Entity Recognition(NER), it was possible to identify and classify the words in a comment. Entities can be people, organisations, dates, locations, numbers. We used the library Spacy to obtain the entities of each comment.
- **Proper Nouns:** A proper noun is an entity specific for person, organisations or places, we used the proper noun to focus the analysis on the name of the players and the clubs. We used the library Spacy to obtain the Proper nouns of each comment.

Associated with the identification of entities and proper nouns it is relevant to understand what is the frequency of this terms on the text. This is a topic very studied in many aspects of linguistic.

The Mandelbrot distribution describes in satisfactory way the probability of the distribution of the terms. In this case, the calculation used to find the distribution is fitted using Maximum Likelihood Estimation (MLE) and is described in Ramos [36].

#### 4.2.1. Global

In this exploratory work it was possible to compare the behaviour of the sentiment analysis among the data sources:

- As showed in Figure 13, there is more dispersion on the results of Stanza analysis of polarity and the result of TextBlob library shows a concentrated polarity on the middle of the chart which indicates that the library interpreted a neutral polarity in the three data sources. We can observe a concentrated presence of

TABLE 7. Comparison of rank and mean across layers and entity sources - Object Related Processes

Key ID	Key phrase	Reddit	Reddit	Live	Live	Ratings	Ratings
		Rank	Mean	Commentary	Commentary	Rank	Mean
L1.62	Recognise when and how to support team members	1	0.410	1	0.374	1	0.330
L1.30	Recognise/anticipate team member actions	2	0.390	2	0.355	2	0.311
L1.15	Initial distribution of the ball	3	0.385	3	0.353	3	0.303
L1.26	Organise team members at opposition set pieces	4	0.379	5	0.342	4	0.294
L1.21	Provide protection from injury	5	0.373	4	0.343	5	0.293
L1.43	Force opposition wide and back	6	0.372	6	0.339	6	0.289
L1.60	Switch field of play	7	0.369	7	0.338	9	0.285
L1.52	Close ball control	8	0.366	8	0.337	10	0.285
L1.27	Provide visual personal identification	9	0.365	9	0.336	8	0.287
L1.29	Enhances physiological performance	10	0.363	10	0.335	7	0.288
L1.3	Increase foot traction	11	0.360	11	0.332	11	0.283
L1.28	Provide team identity	12	0.360	12	0.331	12	0.283
L1.67	Stretch opposition defensive lines	13	0.359	14	0.328	16	0.278
L1.22	Provides match tactics	14	0.358	13	0.330	13	0.282
L1.57	Control speed of game	15	0.354	18	0.325	20	0.271
L1.5	Delay attacks	16	0.353	15	0.328	15	0.279
L1.36	Provide spatial awareness	17	0.353	16	0.327	14	0.282
L1.56	Recognise speed of game	18	0.353	19	0.325	19	0.274
L1.47	Understand role in attack	19	0.352	17	0.325	18	0.275
L1.51	Effective touches forward	20	0.351	20	0.324	17	0.276
L1.69	Get into scoring positions	21	0.348	21	0.318	23	0.266
L1.20	Aerial challenges	22	0.343	22	0.317	22	0.269
L1.31	Recognise/anticipate opposition actions	23	0.342	25	0.313	21	0.270
L1.25	Manage defensive line	24	0.342	23	0.314	25	0.261
L1.19	Reaction time	25	0.339	24	0.313	24	0.264
L1.41	Attack at set pieces	26	0.338	27	0.309	30	0.257
L1.40	Shooting	27	0.336	26	0.310	27	0.259
L1.2	Receiving passes	28	0.334	28	0.308	29	0.258
L1.23	Nonverbal communication	29	0.334	29	0.308	28	0.259
L1.49	Create space for self and team members	30	0.334	35	0.301	38	0.249
L1.38	Provide playing surface	31	0.333	30	0.307	26	0.260
L1.59	Open passing lanes	32	0.329	31	0.304	34	0.252
L1.45	Runs in behind	33	0.327	33	0.302	32	0.254
L1.66	Close passing lanes	34	0.327	34	0.302	33	0.253
L1.55	Pressure opposition	35	0.326	32	0.303	31	0.255
L1.4	Understand role in defence	36	0.324	37	0.300	37	0.251
L1.39	Tackling	37	0.322	36	0.300	36	0.251
L1.48	Drop back	38	0.321	39	0.298	39	0.248
L1.14	Handling	39	0.321	38	0.298	35	0.252
L1.53	Free kicks	40	0.320	40	0.295	43	0.245
L1.54	Secondary ball wins	41	0.319	42	0.294	42	0.245
L1.63	Vision	42	0.318	41	0.294	41	0.245
L1.44	Break lines	43	0.315	43	0.292	46	0.243
L1.17	Footwork	44	0.314	46	0.290	49	0.241
L1.9	1v1	45	0.313	44	0.291	40	0.245
L1.61	Composed in possession	46	0.313	45	0.290	45	0.244
L1.46	Crossing the ball	47	0.312	48	0.287	52	0.237
L1.42	Opposition marking	48	0.310	47	0.290	44	0.244
L1.68	Recognise time and type of runs	49	0.309	53	0.282	54	0.234
L1.37	Provide playing boundaries	50	0.309	49	0.287	47	0.242
L1.18	Cut down angles	51	0.307	50	0.286	51	0.241
L1.50	Comfortable on the ball	52	0.306	54	0.280	59	0.229
L1.33	Prediction	53	0.305	51	0.285	50	0.241
L1.35	Demonstrate respect	54	0.302	52	0.283	48	0.242
L1.10	Long balls	55	0.302	55	0.280	58	0.230
L1.32	Perception	56	0.300	56	0.280	53	0.235
L1.24	Verbal communication	57	0.299	57	0.277	55	0.232
L1.6	Deny attacks	58	0.296	58	0.277	57	0.231
L1.1	Passing	59	0.296	59	0.276	56	0.231
L1.13	Punch	60	0.293	60	0.273	60	0.227
L1.8	Protect the ball	61	0.291	61	0.267	61	0.216
L1.58	Understand and maintain team culture	62	0.276	64	0.252	65	0.204
L1.7	Hold the ball	63	0.274	66	0.251	67	0.200
L1.64	Creativity	64	0.273	62	0.256	62	0.212
L1.12	Tip	65	0.271	63	0.255	63	0.212
L1.16	Ball control and kicking	66	0.270	67	0.244	68	0.193
L1.11	Dive	67	0.266	65	0.251	64	0.206
L1.34	Understand coach's intent	68	0.243	68	0.229	66	0.202
L1.65	Risk-taking	69	0.200	69	0.188	69	0.144

TABLE 8. Comparison of rank and mean across layers and entity sources - Purpose-Related Functions

Key ID	Key phrase	Reddit	Reddit	Live	Live	Ratings	Ratings
		Rank	Mean	Commentary	Commentary	Rank	Mean
L2.7	Maintain position in team structure	1	0.456	1	0.413	1	0.365
L2.15	Play in line with coach ethos	2	0.455	2	0.411	2	0.363
L2.13	Appropriate decision-making	3	0.444	3	0.407	3	0.361
L2.18	Manage own fitness physical condition	4	0.443	5	0.406	5	0.358
L2.12	Maintain resilience	5	0.442	4	0.406	4	0.360
L2.10	Maximise time and space	6	0.437	6	0.402	6	0.355
L2.4	Develop and maintain situation awareness	7	0.435	7	0.398	7	0.355
L2.16	Adaptability	8	0.431	8	0.398	8	0.352
L2.6	Restrict time and space of opposition	9	0.425	9	0.389	9	0.343
L2.17	Tactical fouls	10	0.404	11	0.366	11	0.321
L2.3	Transition	11	0.399	10	0.370	10	0.322
L2.8	Gain possession	12	0.391	12	0.361	12	0.314
L2.9	Maintain possession	13	0.388	13	0.359	13	0.311
L2.14	Manage match tempo	14	0.387	14	0.352	14	0.301
L2.11	Leadership	15	0.378	15	0.348	15	0.300
L2.2	Attack	16	0.375	16	0.346	16	0.297
L2.5	Communication	17	0.370	17	0.343	17	0.296
L2.1	Defend	18	0.344	18	0.320	18	0.271

TABLE 9. Comparison of rank and mean across layers and entity sources - Value & Priority Measures

Key ID	Key phrase	Reddit	Reddit	Live	Live	Ratings	Ratings
		Rank	Mean	Commentary	Commentary	Rank	Mean
L3.32	Goals scored	1	0.493	1	0.446	1	0.398
L3.21	Runs without the ball	2	0.492	2	0.446	3	0.394
L3.13	Effective defensive clearances	3	0.487	5	0.441	5	0.391
L3.3	Goals conceded	4	0.484	7	0.438	7	0.389
L3.10	Effective contests	5	0.484	3	0.445	2	0.397
L3.27	Tackles won	6	0.481	6	0.439	6	0.391
L3.20	Runs with the ball	7	0.480	10	0.435	18	0.382
L3.24	Shots on goal	8	0.480	18	0.433	20	0.381
L3.23	Effective crosses	9	0.479	4	0.441	4	0.394
L3.4	Saves made	10	0.478	8	0.435	16	0.384
L3.15	Block shots	11	0.475	12	0.434	15	0.384
L3.29	Switch the play	12	0.475	17	0.433	19	0.382
L3.31	Press opposition defenders	13	0.475	16	0.433	10	0.386
L3.8	Clearances	14	0.474	9	0.435	8	0.387
L3.30	Press opposition attack	15	0.474	15	0.434	13	0.385
L3.16	Effective passes	16	0.473	13	0.434	11	0.386
L3.25	Successful 1v1	17	0.473	19	0.432	14	0.385
L3.28	Supporting runs	18	0.472	14	0.434	12	0.385
L3.19	Work rate	19	0.472	11	0.434	9	0.387
L3.1	Positioning	20	0.465	20	0.430	17	0.383
L3.14	Opposition offsides	21	0.464	22	0.421	22	0.374
L3.6	Passing accuracy	22	0.464	21	0.426	21	0.379
L3.17	Duels	23	0.457	24	0.416	24	0.368
L3.11	Retreat defence	24	0.457	23	0.418	23	0.370
L3.9	Shot conceded	25	0.455	25	0.415	26	0.364
L3.22	Headers won	26	0.453	28	0.408	29	0.360
L3.2	Coordinate the defence	27	0.452	26	0.413	25	0.365
L3.12	Pressure ball carrier	28	0.446	27	0.409	28	0.360
L3.26	Overlaps	29	0.444	29	0.408	27	0.361
L3.5	Effective passes and throws	30	0.440	31	0.398	31	0.350
L3.18	Block shots and crosses	31	0.436	32	0.394	32	0.346
L3.7	Interceptions	32	0.432	30	0.399	30	0.352

comments very positives, very negatives or very neutrals on Reddit with Stanza Library, when with the other data sources the distribution is more uniform.

- In Figure 13, TextBlob tends to be more neutral and less discriminative in relation of Stanza which is more distributed. 13 where the Stanza library also shows a several dispersion of polarity.

TABLE 10. Comparison of rank and mean across layers and entity sources  
- Functional Purposes

Key ID	Key phrase	Reddit	Reddit	Live	Live	Ratings	Ratings
		Rank	Mean	Commentary	Commentary	Rank	Mean
L4.22	Assist in goal scoring	1	0.470	3	0.426	5	0.376
L4.8	Create goal scoring opportunities	2	0.470	1	0.430	1	0.381
L4.20	Bring others into offensive situations	3	0.468	2	0.429	2	0.378
L4.7	Break up opposition attacks	4	0.461	5	0.423	4	0.376
L4.3	Provide a safe passing option	5	0.461	6	0.423	6	0.375
L4.19	Provide an outlet	6	0.460	4	0.424	3	0.377
L4.15	Bring attacking players into play	7	0.459	7	0.416	7	0.364
L4.6	Prevent the opposition from scoring	8	0.452	8	0.410	9	0.362
L4.17	Provide attacking support	9	0.447	9	0.410	8	0.362
L4.5	Organise and coordinate team members	10	0.445	10	0.406	10	0.360
L4.9	Prevent attempts at goal and crosses	11	0.443	11	0.400	11	0.351
L4.1	Prevent goals scored	12	0.435	12	0.395	12	0.347
L4.13	Disturb build-up of opposition	13	0.432	13	0.393	13	0.347
L4.16	Score goals	14	0.430	14	0.391	15	0.340
L4.14	Protect central defence	15	0.424	15	0.388	14	0.341
L4.21	Initiate disturb build-up of opposition	16	0.423	16	0.385	16	0.339
L4.11	Connect defence and attacking players	17	0.422	17	0.379	17	0.332
L4.18	Create chances	18	0.409	18	0.378	18	0.330
L4.4	Relieve pressure	19	0.407	19	0.377	19	0.329
L4.12	Assist and continue build-up	20	0.402	21	0.364	20	0.322
L4.2	Initiate build-up	21	0.394	22	0.359	22	0.311
L4.10	Stretch opposition	22	0.393	20	0.364	21	0.318

- In Figure 14, we can observe a concentration of very objective comments on Reddit and a little concentration of Subjective comments on Sports Sites Live Comments data source, which was expected because the type of comments are related with the match events and usually do not express any opinion about the match or the players.
- In Figure 14, for all the media the level of subjectivity is relatively paradoxical, and the subjectivity of Reddit was not expected because there are very objective comments, possibly there is a relation with this result and the Textblob performance
- Generally, we can observe that the Chelsea’s player Kanté was elected the player of the match on the studied match, but on the Sports Sites the most valuable player was Chelsea’s player Kai Havertz, who scored the winner goal of the match. Kanté was one of the Top 10 commented topics just on the Reddit, but Kai Havertz appears on the Top 10 commented topics of all the three data sources.
- In the other hand, Chelsea’s player Timo Werner appears on the Top 10 commented topics of all the three data sources, but his performance was very criticised in the Sports Sites Performance Analysis, and the polarity of the comments about Timo Werner can attest, he was very commented with very negative comments as illustrated on the polarity charts.
- Chelsea, the winner team, was the most commented entity/proper noun in all the datasets.
- Figure 15 shows a great dispersion of the objective performance among the players, which can have a relation with Figure

- In the sample in Figure 16, it is possible to perceive that the results, in some comments, the two libraries (Stanza and TextBlob) are very similar, but in the third comment in the figure, Stanza calculated a negative polarity of -1.0 while TextBlob computed 0.0, and when we analyse the entry the conclusion is that this is a negative sentence, which reveal the problems of using a generic tool in a specific domain such as sports.

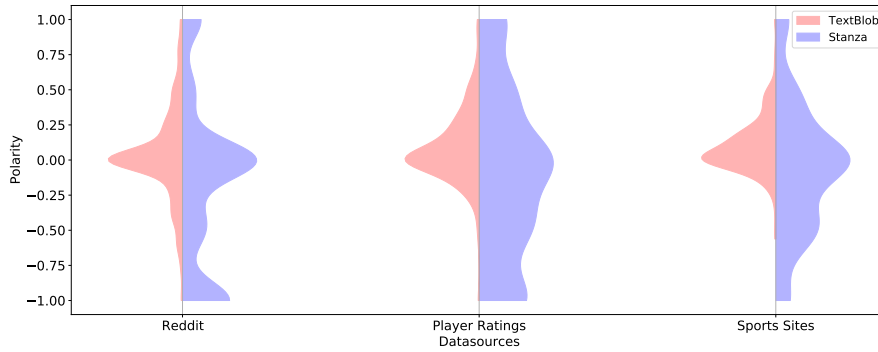


FIGURE 13. Global Polarity TextBlob/Stanza

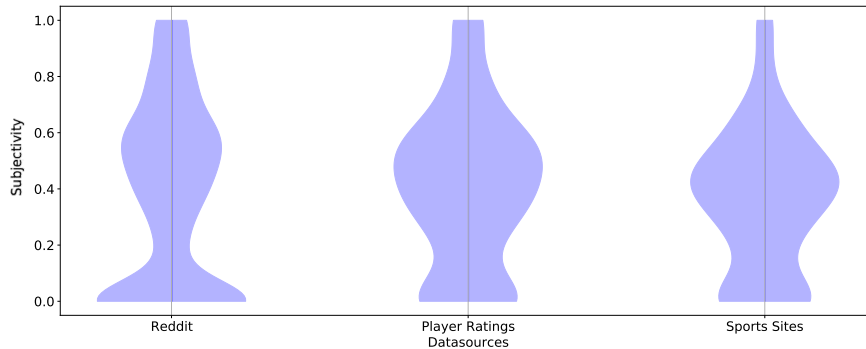


FIGURE 14. Global Subjectivity TextBlob

#### 4.2.2. Reddit

In the analysed 1791 comments (spanning 5 threads), we have found 705 different entities and 721 different proper nouns. The top 3 of the commented entities related with the Clubs, Players and Managers were: “Chelsea”, “Werner”, and “Kante” and the top 3 of the commented proper nouns related with the Clubs were: “Chelsea”, “Pep”, and “City” .

- “Tuchel”, Chelsea’s coach, was among the top 10 most commented proper nouns, had the most positive polarity detected by TextBlob library followed by “Kante”, the player elected the man of the match.
- “Kai” and “Havertz” refers to the same player but the polarity presented a different result for each proper noun.

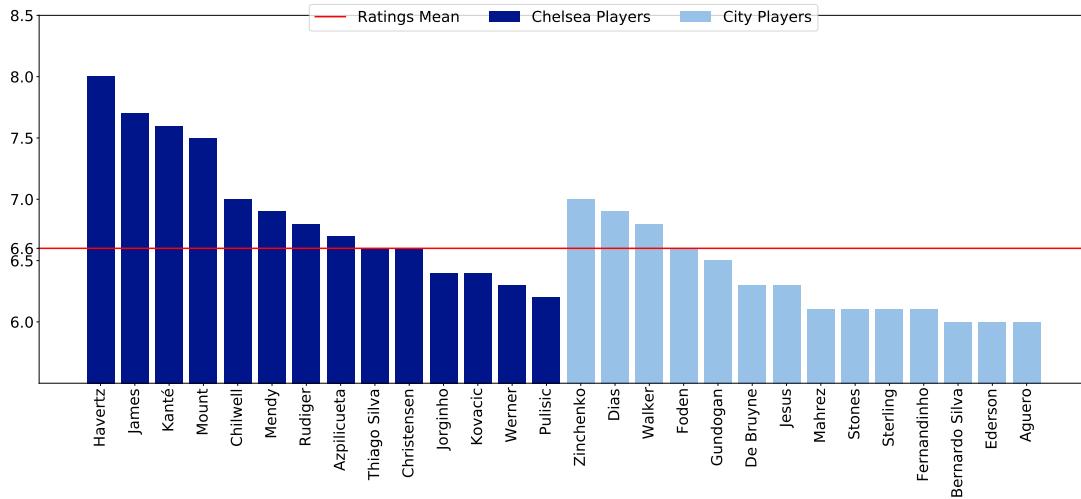


FIGURE 15. Players Ratings (WhoScored.com) - UEFA Champions League Final

Stanza Polarity: -0.5

TextBlob Polarity/Subjectivity: Sentiment(polarity=-0.4, subjectivity=0.6)

pimpsquadforlife

Terrible from Zinny, stones, and Dias.

Zinny 14 19 PERSON

Stanza Polarity: -1.0

TextBlob Polarity/Subjectivity: Sentiment(polarity=-1.0, subjectivity=1.0)

narzivial

Remind me why sterling is on again?

Stanza Polarity: -1.0

TextBlob Polarity/Subjectivity: Sentiment(polarity=0.0, subjectivity=0.0)

FlatlineMonday

why in the world is Gundogan playing as a defensive  
midfielder?

All our problems stem from that.

I am so frustrated.

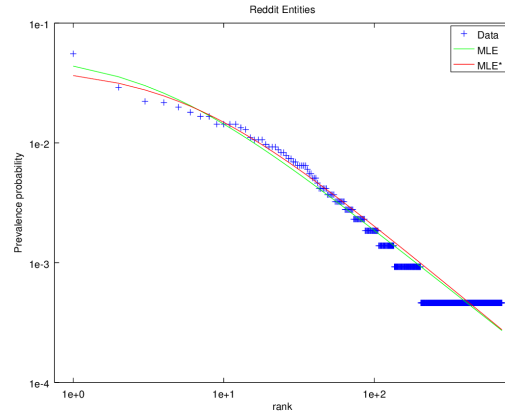
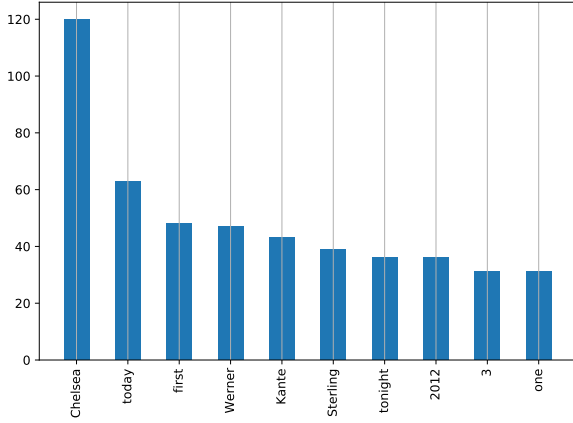
Gundogan 20 28 PERSON

FIGURE 16. Sports Site Live Comments Example

Figure 17 shows the top 10 most commented entities, the quantity of times that they were mentioned, and the corresponding Mandelbrot distribution probabilities fitted using Maximum Likelihood Estimation (MLE) of the studied dataset.

Figure 18 shows the top 10 most commented proper nouns and the quantity of times that they were mentioned, and the corresponding Mandelbrot distribution probabilities fitted using Maximum Likelihood Estimation (MLE) of the studied dataset.

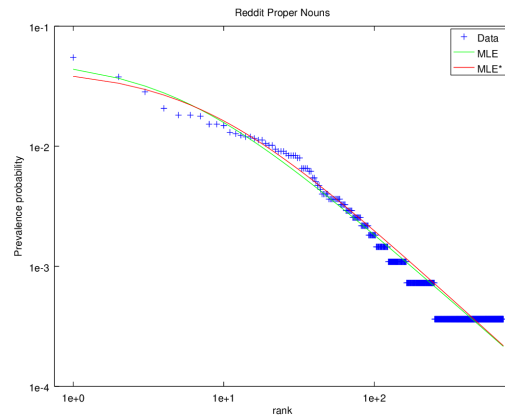
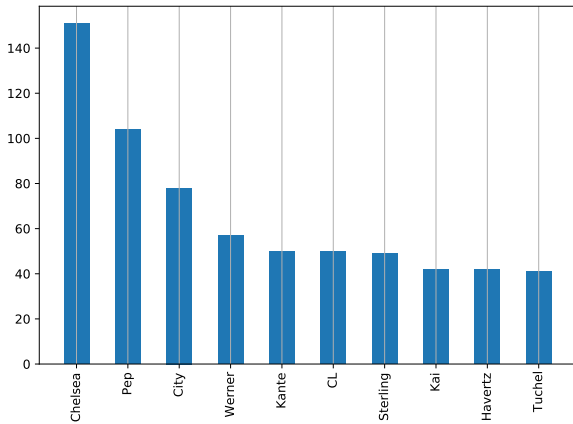




(A) Top 10 commented Entities - Reddit

(B) Prevalence probability Entities - Reddit

FIGURE 17. Top 10 commented Entities and MLE Analysis - Reddit



(A) Top 10 commented Proper Nouns - Reddit

(B) Prevalence probability Proper Nouns - Reddit

FIGURE 18. Top 10 commented Proper Nouns and MLE Analysis - Reddit

In Figures 17 and 18, the number of times each term was cited follows the Mandelbrot Distribution with the exception of the first term “Chelsea”, the Winner team. The detected entities correspond to words not related with the match and players performance (e.g., “today”, “first”).

Figures 19, 20, 21, 22, 23, 24, show the comparison between the top 10 entities and proper nouns polarity/ subjectivity with the global polarity/subjectivity in the related data source.

In Figures 19 and 20, for the entities directly related with the match the polarity tends to be more positive and the terms related with the winner team tend to be more positive.

In Figures 21 and 22, With exception of “Kante” and “Havertz” (Man of the match and the author of winning goal, respectively) the subjectivity is not much different than the global subjectivity, the comments tend to be more objective than subjective.

In Figures 23 and 24, in relation of teams, Stanza did not detect a polarity much negative, but with the teams elements(e.g., players and coaches) there is a great difference of polarity between ‘‘Tuchel’’(positive) Chelsea’s coach and ‘‘Pep’’(negative) City’s Coach and ‘‘Kanté’’, Chelsea’s player (positive) and ‘‘Sterling’’ (negative) City’s player, for example.

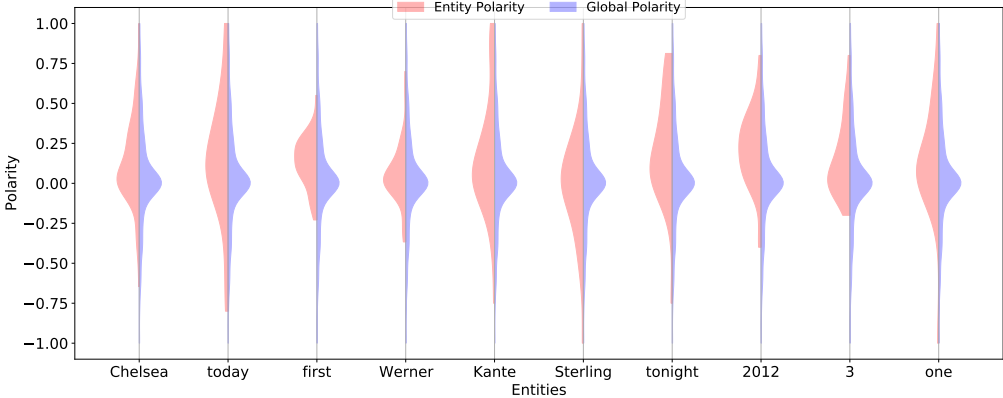


FIGURE 19. Entity Polarity TextBlob - Reddit

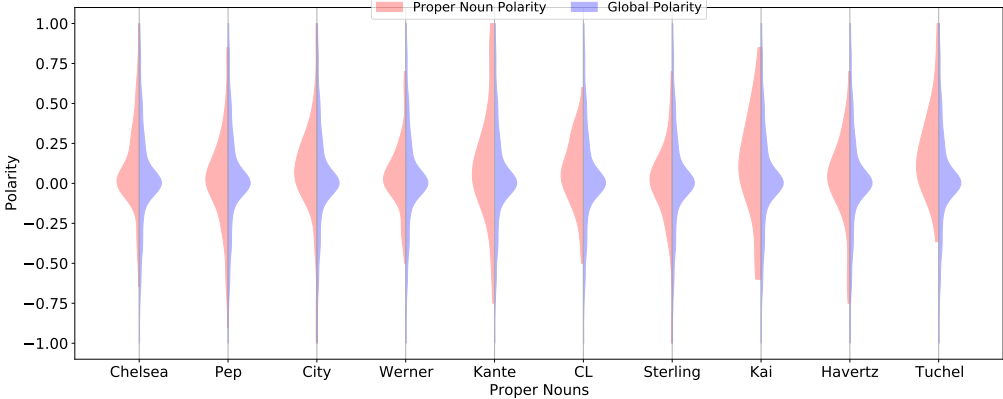


FIGURE 20. Proper Nouns Polarity TextBlob - Reddit

**4.2.3. Player Ratings**

In the analysed 331 comments (spanning 19 sites), we have found 334 different entities and 180 different Proper Nouns. The top 3 of the commented entities related with the Clubs, Players and Managers were: ‘‘Chelsea’’, ‘‘Harvertz’’ and ‘‘Werner’’ and the top 3 of the commented proper nouns related with the Clubs were: ‘‘Chelsea’’, ‘‘City’’, and ‘‘Havertz’’.

- ‘‘Havertz’’, the player who made the winning goal had the most positive polarity detected

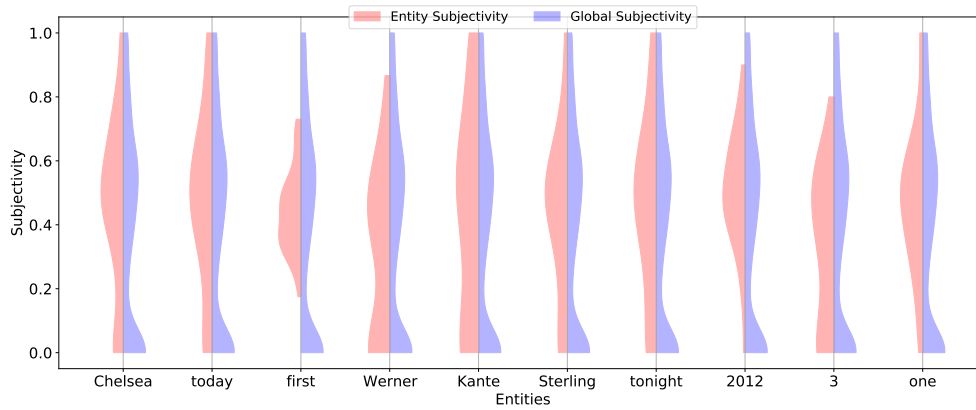


FIGURE 21. Entity Subjectivity TextBlob - Reddit

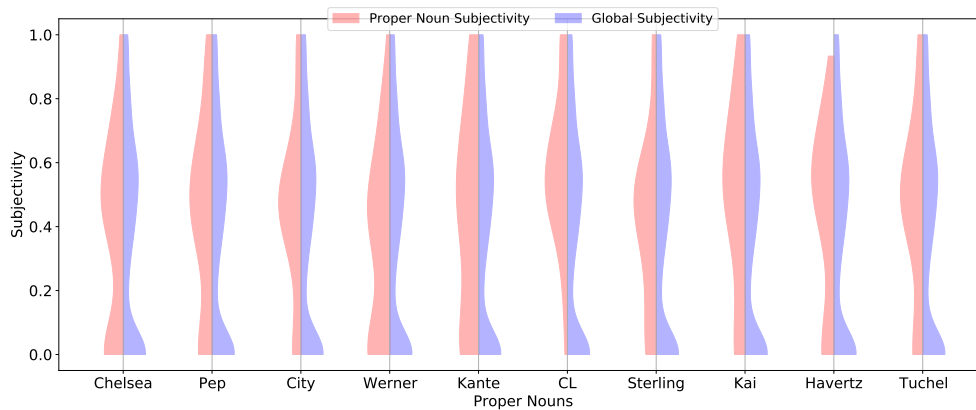


FIGURE 22. Proper Nouns Subjectivity TextBlob - Reddit

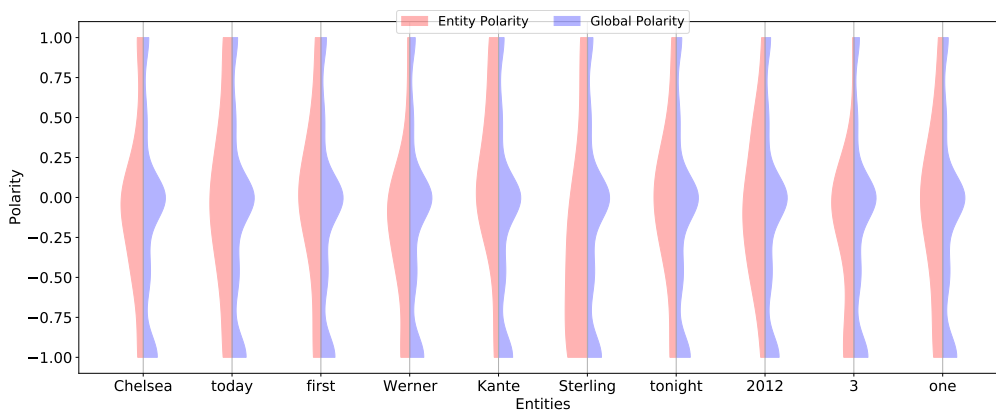


FIGURE 23. Entity Polarity Stanza - Reddit

- “Werner” even playing on the winner team, had a negative polarity detected in the formal media(Player Ratings) perspective.

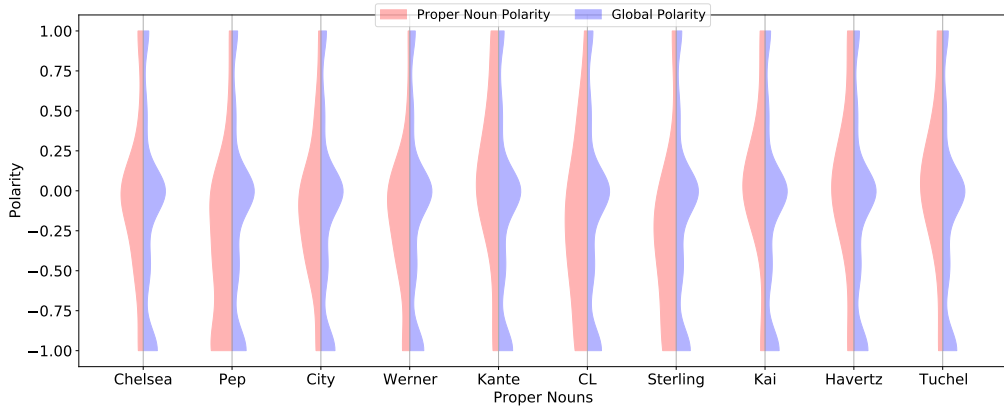
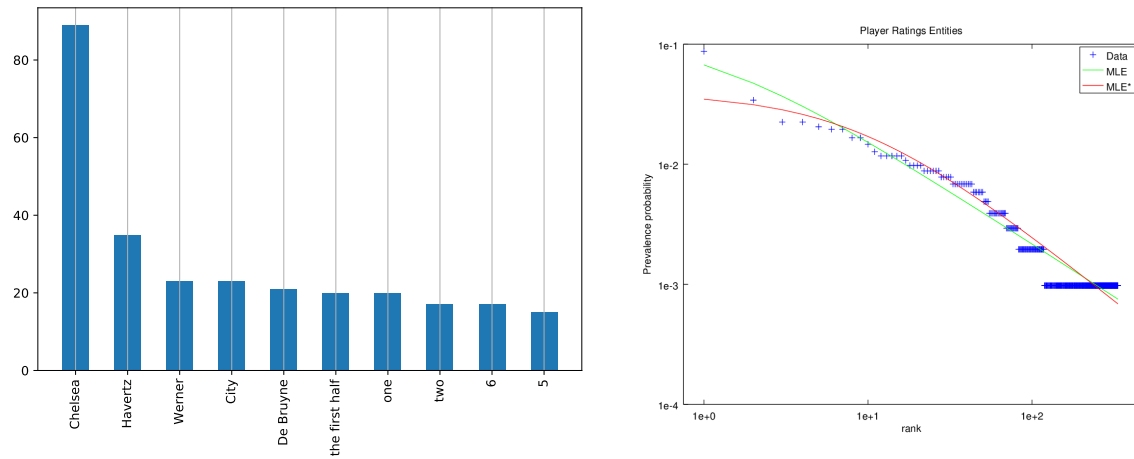


FIGURE 24. Proper Nouns Polarity Stanza - Reddit

- The Figure 25 contains the number “5” and “6” in the top 10 commented entities, probably the two most given ratings, what is befitting with Figure 15 that shows the rating of the players on the match.
- The curve in Figure 26 has a more equalised distribution, because each player is rated once, what means that the count of times each player appears on this dataset is very close.

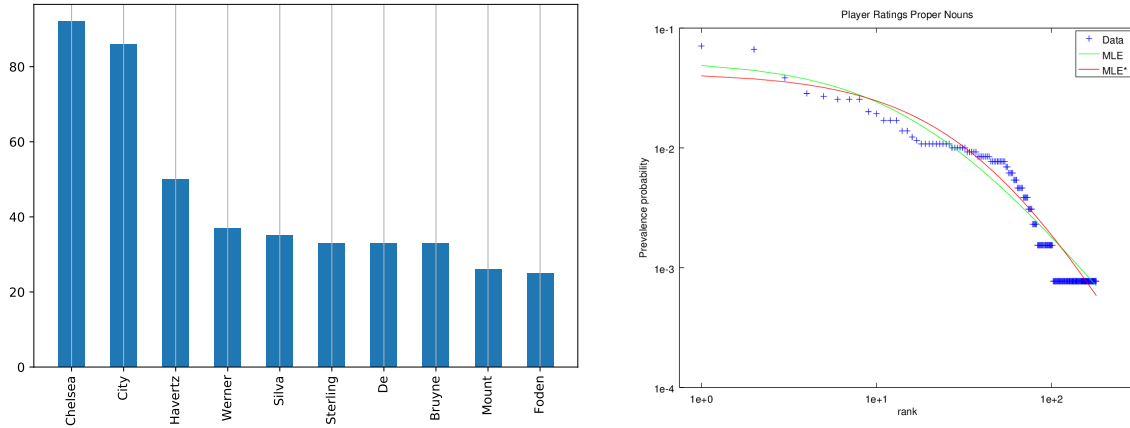
Figure 25 shows the top 10 most commented entities and the quantity of times that they were mentioned, and the corresponding Mandelbrot distribution probabilities fitted using Maximum Likelihood Estimation (MLE) of the studied dataset.



(A) Top 10 commented Entities - Player Ratings (B) Prevalence probability Entities - Player Ratings

FIGURE 25. Top 10 commented Entities and MLE Analysis - Player Ratings

Figure 26 shows the top 10 most commented entities and the quantity of times that they were mentioned, and the corresponding Mandelbrot distribution probabilities fitted using Maximum Likelihood Estimation (MLE) of the studied dataset.



(A) Top 10 commented Proper Nouns - Player Ratings (B) Prevalence probability Proper Nouns - Player Ratings

FIGURE 26. Top 10 commented Proper Nouns and MLE Analysis - Player Ratings

In Figures 25 and 26, there are entities that does not have directly relation with the teams (e.g., “one”, “two”).

The terms does not follow the Mandelbrot distribution, this occurs because in this context there is just one entry per player, what means that the quantity of entities are very similar.

Figures 27, 28, 29, 30, 31, 32 show the comparison between the Top 10 entities and proper nouns polarity/ subjectivity with the global polarity/subjectivity in the related data source.

In Figures 27 and 28, it is possible to distinguish the player of the winner team and the players of the losing team: “Havertz”, “Mount” have a very positive polarity in contrast with “De Bruyne” and “Silva” with a polarity more negative.

In Figures 29 and 30, curiously the distribution is very similar with the global mean and there is no distinction between the teams. The opinion about “Sterling” and “Silva” are very subjective both players with negative polarity too.

In Figures 31 and 32, “Werner” did not achieve any comment polarity with rate greater than 0.5, what may be reflects his participation on the match, even playing on the winner team.

#### 4.2.4. Sports Sites Comments

In the analysed 257 comments (spanning 5 sites), we have found 535 different entities and 384 different Proper Nouns. The top 3 of the commented entities related with the Clubs, Players and Managers were: “Chelsea”, “first” and “Man City” and the top 3 of the commented proper nouns related with the Clubs were: “Chelsea”, “City” and “League” .

- “Havertz” player who made the winner goal had the most positive polarity detected;

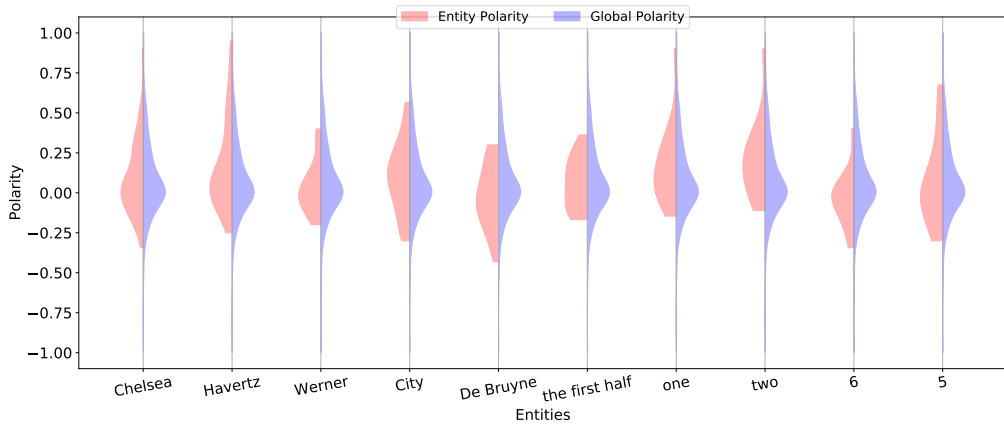


FIGURE 27. Entity Polarity TextBlob - Players Ratings

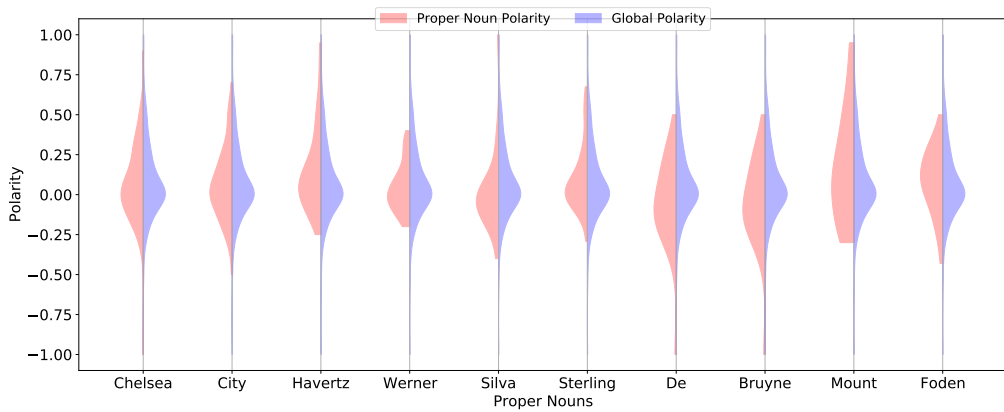


FIGURE 28. Proper Nouns Polarity TextBlob - Players Ratings

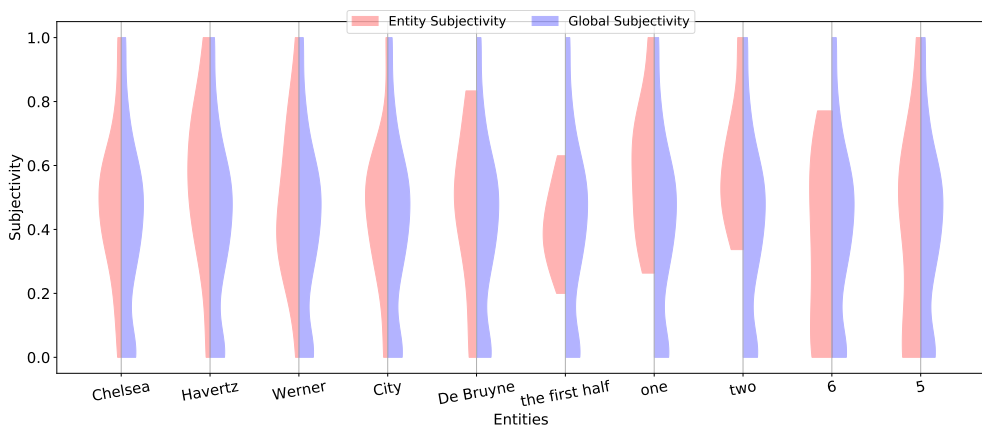


FIGURE 29. Entity Subjectivity TextBlob - Players Ratings

- “Werner” even playing on the winner team, had a negative polarity detected by the formal media(Sports Sites Comments) perspective;

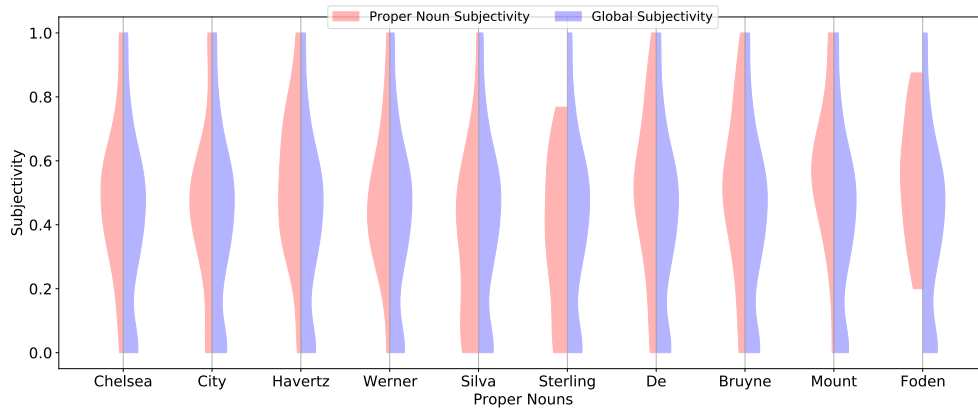


FIGURE 30. Proper Nouns Subjectivity TextBlob - Players Ratings

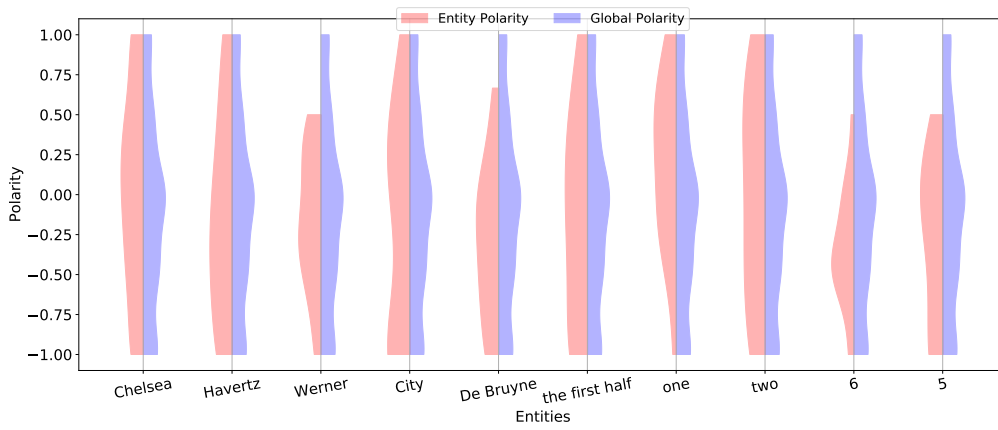


FIGURE 31. Entity Polarity Stanza - Players Ratings

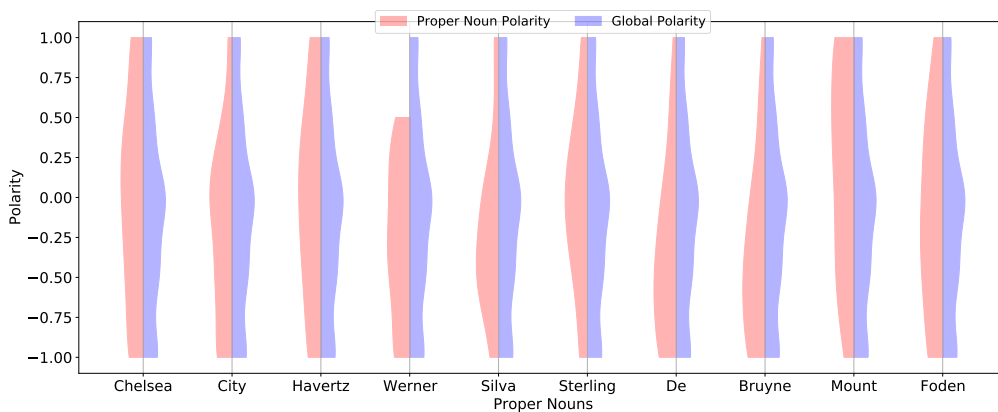
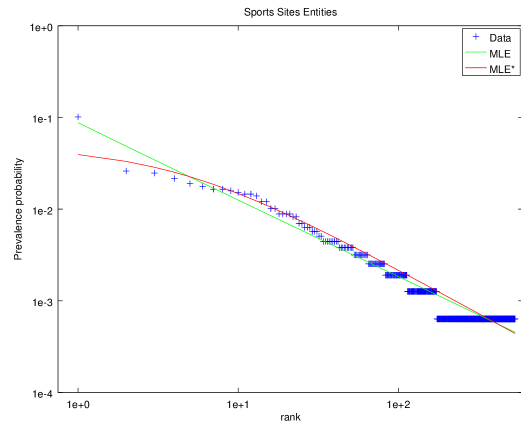
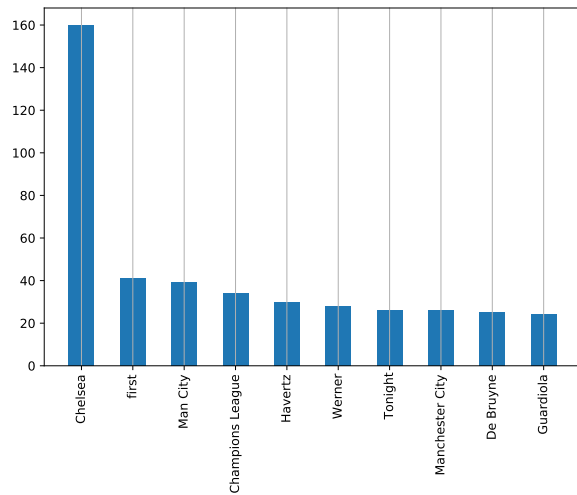


FIGURE 32. Proper Nouns Polarity Stanza - Players Ratings

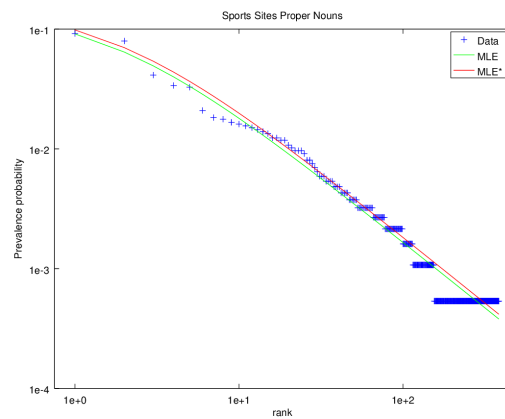
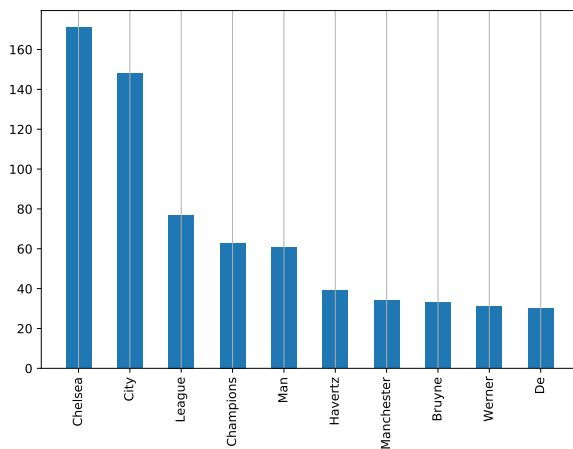
Figure 33 shows the top 10 most commented entities and the quantity of times that they were mentioned.



(A) Top 10 commented Entities - Sports Sites Comments (B) Prevalence probability Entities - Sports Sites Comments

FIGURE 33. Top 10 commented Entities and MLE Analysis - Sports Sites

Figure 34 shows the top 10 most commented entities and the quantity of times that they were mentioned, and the corresponding Mandelbrot distribution probabilities fitted using Maximum Likelihood Estimation (MLE) of the studied dataset.



(A) Top 10 commented Proper Nouns - Sports Sites Comments (B) Prevalence probability Proper Nouns - Sports Sites Comments

FIGURE 34. Top 10 commented Proper Nouns and MLE Analysis - Sports Sites

In Figures 33 and 34, there is again a relation with the commented terms and the Mandelbrot Distribution.

Figures 35, 36, 37, 38, 39, 40, show the comparison between the top 10 entities and proper nouns polarity/subjectivity with the global polarity/subjectivity in the related data source.

In Figure 35 and 36, in general, the polarity is less extreme than Reddit, and it is natural that fans have more extreme comments than specialised media.



In Figures 37 and 38, The comments are more objectives than the other sources, what make senses because the content of these comments are totally related with the match events.

In Figures 39 and 40, polarity tends to be more selective, for example, there are no comments very negatives about “Werner” neither comments very positives about “De Bruyne”. “Man City”,the losing team, has the greatest dispersion of polarity.

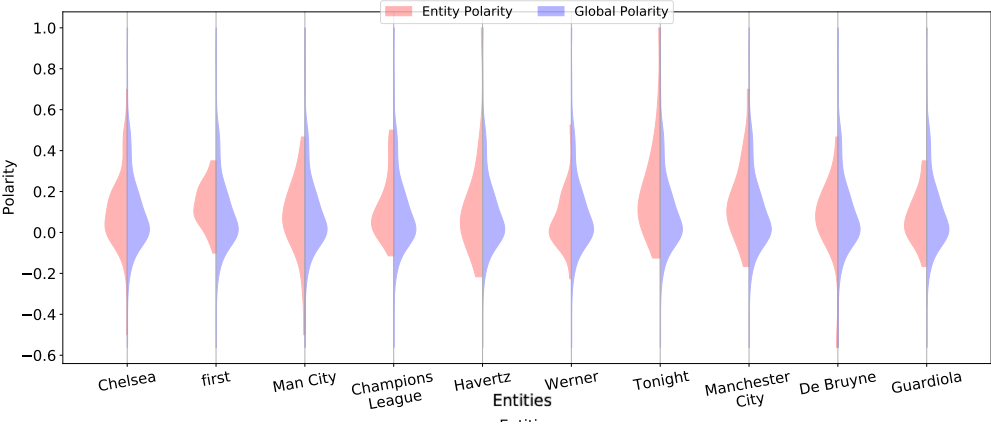


FIGURE 35. Entity Polarity TextBlob - Sports Sites Comments

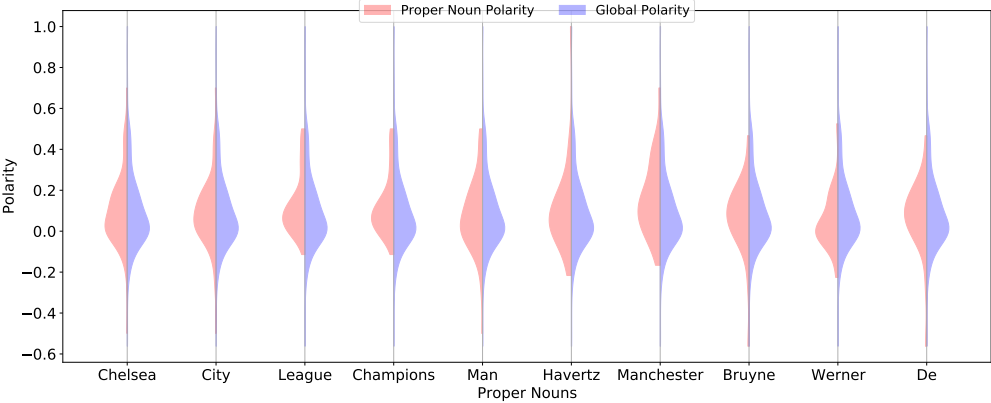


FIGURE 36. Proper Nouns Polarity TextBlob - Sports Sites Comments

It is important to mention that the “Chelsea” team is commonly know as “Blues” and “Manchester City” is commonly know as “Cityzens”, these are examples to show that we are not clustering all references to the same entity .

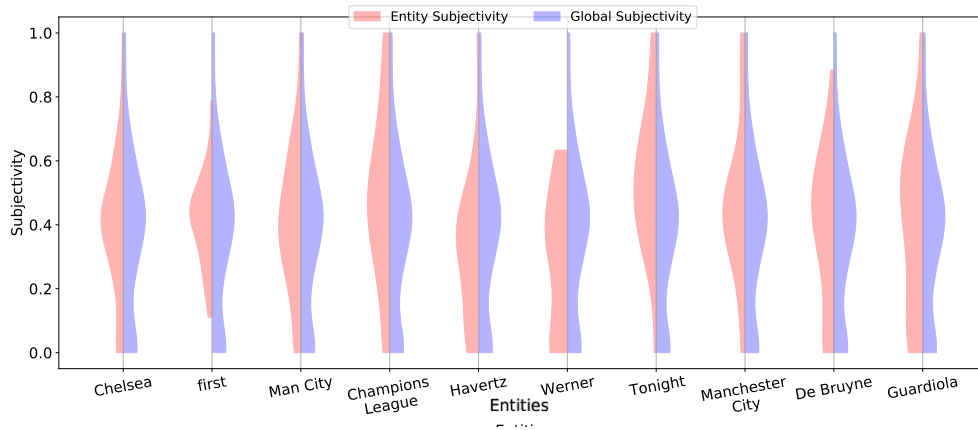


FIGURE 37. Entity Subjectivity TextBlob - Sports Sites Comments

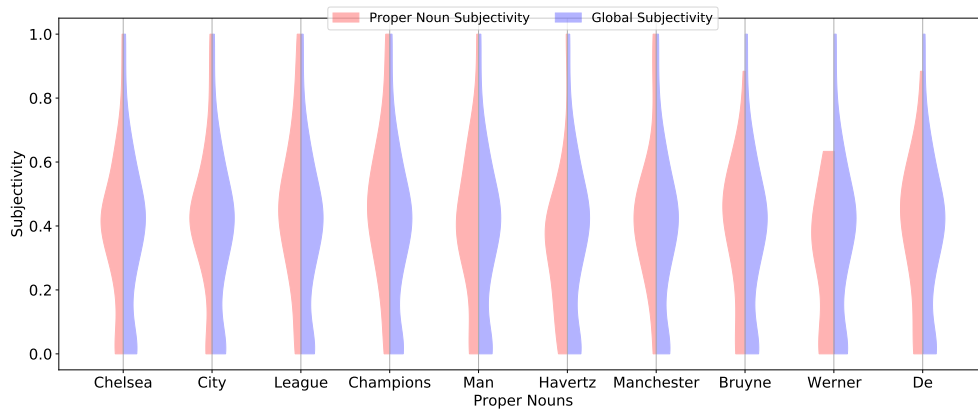


FIGURE 38. Proper Nouns Subjectivity TextBlob - Sports Sites Comments

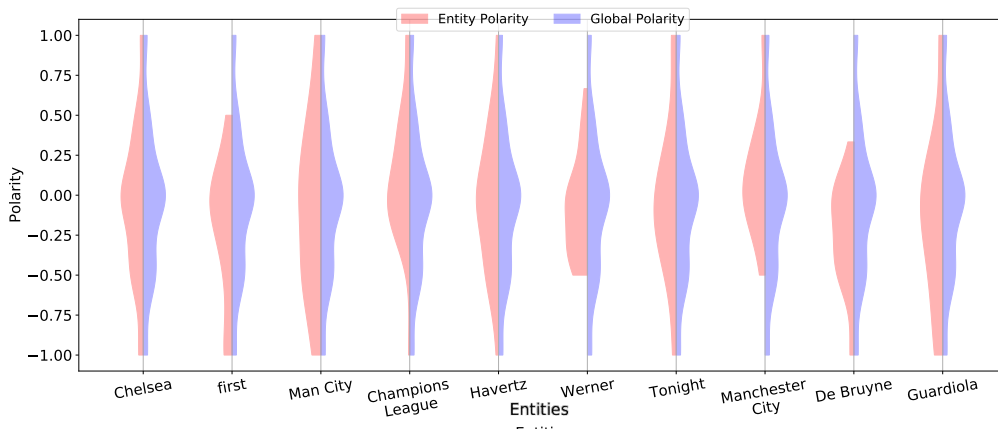


FIGURE 39. Entity Polarity Stanza - Sports Sites Comments

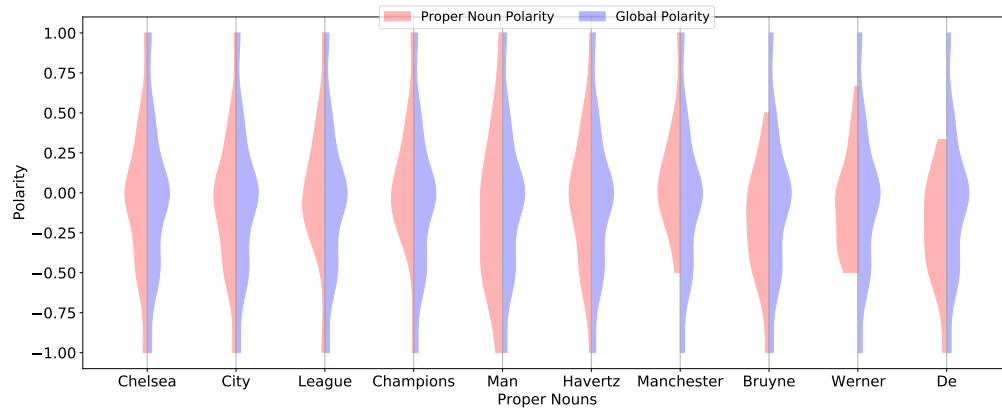


FIGURE 40. Proper Nouns Polarity Stanza - Sports Sites Comments



## Conclusions and Future Research

### 5.1. Global Perspective

In a global perspective, it was possible to obtain common conclusions of both exploratory works:

- The behaviour of the libraries can be improved if adapted to Football context, for example, “Manchester” is a city in England, but in Football Context this entity corresponds more of a two of the biggest teams of English football league than the city.
- The software does not differentiate players with the same name or surname, in the case of study we had two occurrences, “Bernardo Silva” Manchester City’s player and “Thiago Silva” Chelsea’s player and “Benjamin Mendy” Manchester City’s player and “Édouard Mendy” Chelsea’s player.

Concerning future work, we have some ideas to increase the accuracy of the results:

- Create a repository with more data from more sources (Social Media and Formal Media) adapted to Football Context;
- Try to adapt the used methods to other sports;
- Try to use other methods of Sentiment and Semantic Analysis to compare with current results;
- Use more matches as case of study to find behaviours and increase the accuracy.
- The creation of a specific platform to connect football fans and the Data Department of the Football Teams could lead to an integrated (qualitative and quantitative) perspective on performance analysis.

### 5.2. Semantic Similarity

In the work described in this paper it was possible to explore how key phrases associated to different levels of Work Domain Analysis are used in football matches commentary published electronically by different sources. From this exploratory work the following conclusions, could be obtained:

- The similarity score between commentary entries and WDA key phrases shows a great dispersion across all domains (sources, levels, and entries);
- The higher similarity values are obtained at the WDA level *L3*. *Values & priority measures*. It is worth of note that the key phrases identified at this level have usually a closely related match annotation item (e.g., Goals scored);

- Contrary to what may be expected, comments from users in social media, such as Reddit, present, at all WDA levels, higher semantic similarity values than commentary entries in formal media.

Concerning future work, we foresee six main ideas on how to increase the potential of this project:

- The informal and formal media have close similarity scores, with higher similarity values being achieved by fans comments – it is important to understand how this conclusion generalise to other matches;
- Perform a more comprehensive study of the different key phrases, notably their relative ranking and their potentially hierarchical structure (e.g., Goals and Goals scored/conceded or Runs and Runs with/without the ball).
- The polarity of the sentiment of the fans perspective can provide unanticipated insights concerning performance analysis of football players. Sentiment analysis captures the subjective part of the football performances, and performance analysis based on metrics (stats about passes, goals, and assists, for example) is the objective part;
- Apply our method to other social media platforms and other sources of formal media commentary notably, try to compare the users behaviour in different platforms;

### 5.3. Sentiment Analysis

There are some impressions that can be studied in the future to improve the accuracy of the results:

- The most of the positive comments followed the winner team, it is necessary to analyse more matches to verify if it is a pattern or there are other factors to determine the most positive commented team.
- Based on the comments of the users, Reddit threads looked like very objective, with specific comments about match events and factual information rather than comments to depreciate or argue about topics outside the focus of the thread.
- The ranking of the most commented entities/proper nouns follows the players participation on the match.

Comments about football use much irony and the tool interprets some positive comments as negative, and vice versa. Another limitation in this work are the nicknames and slang to refer to players, teams and managers, for example, “Azpilicueta”, “Azpi” and “Cesar Azpilicueta” are words to refer to the same player, but the library does not understand that these different entities can refer to the same player. Another limitation found was the perception with entities with more than one meaning, for example, “Manchester” is related with a location, but also is related with a football team. In this context “Manchester” is much more referred as a team than as a location, but the used tools did not interpret it that way. It is necessary to understand, that social media world

is very wide, there is no obligation to respect grammatical rules and write everything in the right way, using the example above, “azplicueta” also refers to the same player but was written in an incorrect way.

Concerning future work, there are five main ideas on how to increase the potential of this project:

- Verify the relation between the performance analysis of the football players and the sentiment analysis of the fans perspective. Sentiment analysis is the subjective part of the football perspective, and the performance analysis based on metrics is the objective one, where we can easily assess players’ performance based on their stats, e.g., passes, goals, and assists.
- Try to adapt the used methods to the sports domain. As show in Figure 5, the two main goals of football teams are attracting traffic to website and communicate with fans, through the platform this could be done better, focusing on a target audience.
- Analyse the data with others machine learning tools and/or create a mechanism to estimate the polarity, comparing the performance of the sentiment analyses approaches.
- Apply our method to other social media platforms and try to compare the users behaviour for each platform.





## References

- [1] F. Soriano, *A bola não entra por acaso: estratégias inovadoras de gestão inspiradas no mundo do futebol*. Larousse do Brasil, 2010, ISBN: 9788576356974.
- [2] S. Dobson and J. Goddard, *The Economics of Football*, 2nd ed. Cambridge University Press, 2011. DOI: 10.1017/CB09780511973864.
- [3] R. Cotta, *Análise de desempenho no futebol: Entre a teoria e a prática*. 1st ed. Appris, 2018, ISBN: 9788547315740.
- [4] M. Vergeer and L. Mulder, “Football players’ popularity on twitter explained: Performance on the pitch or performance on twitter?” *International Journal of Sport Communication*, vol. 12, pp. 376–396, Sep. 2019. DOI: 10.1123/ijsc.2018-0171.
- [5] K. Ballouli, “It’s a Whole New Ballgame: How Social Media is Changing Sports,” *Sport Management Review*, vol. 15, pp. 381–382, Aug. 2012. DOI: 10.1016/j.smr.2012.02.008.
- [6] A. M. W. C. Carling and T. Reilly, *Handbook of soccer match analysis: A systematic approach to improving performance*, 1st. Abingdon, UK: Routledge, 2005.
- [7] E. Rampinini, A. Coutts, C. Castagna, R. Sassi, and F. Impellizzeri, “Variation in top level soccer match performance,” *International journal of sports medicine*, vol. 28, pp. 1018–24, Dec. 2007. DOI: 10.1055/s-2007-965158.
- [8] E. Rampinini, F. Impellizzeri, C. Castagna, A. Coutts, and U. Wisloff, “Technical performance during soccer matches of the italian serie a league,” *Journal of science and medicine in sport / Sports Medicine Australia*, vol. 12, pp. 227–33, Dec. 2007. DOI: 10.1016/j.jsams.2007.10.002.
- [9] T. Reilly, B. Drust, and N. Clarke, “Muscle fatigue during football match-play,” *Sports medicine (Auckland, N.Z.)*, vol. 38, pp. 357–67, Feb. 2008. DOI: 10.2165/00007256-200838050-00001.
- [10] J. Ekstrand, M. Hägglund, and M. Waldén, “Injury incidence and injury patterns in professional football: The uefa injury study,” *British journal of sports medicine*, vol. 45, pp. 553–8, Jun. 2009. DOI: 10.1136/bjism.2009.060582.
- [11] V. Gouttebauge and G. Kerkhoffs, “Mental health in professional football players,” in Mar. 2018, pp. 851–859, ISBN: 978-3-662-55712-9. DOI: 10.1007/978-3-662-55713-6\_65.
- [12] M. Zacharko, M. Konefał, P. Chmura, J. Baranowski, M. Andrzejewski, K. Blazejczyk, and J. Chmura, “The influence of thermal stress on physical and tactical activities of football players - the lesson from russia’2018 world cup,” Jun. 2019.

- [13] N. Cecchi, D. Monroe, W. Moscoso, J. Hicks, and D. Reinkensmeyer, “Effects of soccer ball inflation pressure and velocity on peak linear and rotational accelerations of ball-to-head impacts,” *Sports Engineering*, vol. 23, p. 16, Sep. 2020. DOI: 10.1007/s12283-020-00331-0.
- [14] H. Andersson, B. Ekblom, and P. Krstrup, “Elite football on artificial turf versus natural grass: Movement patterns, technical standards, and player impressions,” *Journal of sports sciences*, vol. 26, pp. 113–22, Feb. 2008. DOI: 10.1080/02640410701422076.
- [15] J. Gutiérrez Macías, J. Castellano, D. Casamichana, and J. Sánchez, “Effect of pitch size and time of the match in the physical performance of teams the spanish second division (in spanish),” *Retos: nuevas tendencias en educación física, deporte y recreación*, Aug. 2017.
- [16] B. Levine, J. Stray-Gundersen, and R. Mehta, “Effect of altitude on football performance,” *Scandinavian journal of medicine & science in sports*, vol. 18 Suppl 1, pp. 76–84, Sep. 2008. DOI: 10.1111/j.1600-0838.2008.00835.x.
- [17] H. Liu, M.-A. Gómez, B. Gonçalves, and J. Sampaio, “Technical performance and match-to-match variation in elite football teams,” *Journal of Sports Sciences*, vol. 34, no. 6, pp. 509–518, 2016, PMID: 26613399. DOI: 10.1080/02640414.2015.1117121. eprint: <https://doi.org/10.1080/02640414.2015.1117121>. [Online]. Available: <https://doi.org/10.1080/02640414.2015.1117121>.
- [18] M. D. Bush, D. T. Archer, R. Hogg, and P. S. Bradley, “Factors influencing physical and technical variability in the english premier league,” *International Journal of Sports Physiology and Performance*, vol. 10, no. 7, pp. 865–872, 2015. DOI: 10.1123/ijsp.2014-0484. [Online]. Available: <https://journals.humankinetics.com/view/journals/ijsp/10/7/article-p865.xml>.
- [19] S. Mclean, P. Salmon, A. Gorman, G. Read, and C. Solomon, “What’s in a game? A systems approach to enhancing performance analysis in football,” *PLOS ONE*, vol. 12, e0172565, Feb. 2017. DOI: 10.1371/journal.pone.0172565.
- [20] B. Barros, C. Serrão, and R. Lopes, “Distributed crowd-based annotation of soccer games using mobile devices,” in *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support - Volume 1: icSPORTS*, INSTICC, SciTePress, 2018, pp. 40–48, ISBN: 978-989-758-325-4. DOI: 10.5220/0006927000400048.
- [21] M. Stein, H. Janetzko, T. Breitreutz, D. Seebacher, T. Schreck, M. Grossniklaus, I. Couzin, and D. A. Keim, “Director’s cut: Analysis and annotation of soccer matches,” *IEEE Computer Graphics and Applications*, vol. 36, no. 5, pp. 50–60, Sep. 2016, Special Issue Sports Data Visualization. DOI: 10.1109/MCG.2016.102.
- [22] E. Berber, S. McLean, V. Beanland, G. J. M. Read, and P. M. Salmon, “Defining the attributes for specific playing positions in football match-play: A complex systems approach,” *Journal of Sports Sciences*, vol. 38, no. 11–12, pp. 1248–1258, 2020. DOI: 10.1080/02640414.2020.1768636. eprint: <https://doi.org/10.1080/02640414>.

- 2020.1768636. [Online]. Available: <https://doi.org/10.1080/02640414.2020.1768636>.
- [23] J. McCarthy, J. Rowley, C. Ashworth, and E. Pioch, “Managing brand presence through social media: The case of uk football clubs,” *Internet Research: Electronic Networking Applications and Policy*, vol. 24, Apr. 2014. DOI: 10.1108/IntR-08-2012-0154.
- [24] J. Price, N. Farrington, and L. Hall, “Changing the game? the impact of twitter on relationships between football clubs, supporters and the sports media,” *Soccer and Society*, vol. 14, Jul. 2013. DOI: 10.1080/14660970.2013.810431.
- [25] M. Pronschinske, M. Groza, and M. Walker, “Attracting facebook ‘fans’: The importance of authenticity and engagement as a social networking strategy for professional sport teams,” *Sport Marketing Quarterly*, vol. 21, pp. 221–231, Jan. 2012.
- [26] S. Aloufi and A. E. Saddik, “Sentiment identification in football-specific tweets,” *IEEE Access*, vol. 6, pp. 78 609–78 621, 2018. DOI: 10.1109/ACCESS.2018.2885117.
- [27] D. Chandrasekaran and V. Mago, “Evolution of semantic similarity—a survey,” *ACM Comput. Surv.*, vol. 54, no. 2, Feb. 2021, ISSN: 0360-0300. DOI: 10.1145/3440755. [Online]. Available: <https://doi.org/10.1145/3440755>.
- [28] A. Collins and E. Loftus, “A spreading activation theory of semantic processing,” *Psychological Review*, vol. 82, pp. 407–428, Nov. 1975. DOI: 10.1037//0033-295X.82.6.407.
- [29] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, “A survey of sentiment analysis in social media,” *Knowledge and Information Systems*, vol. 60, Aug. 2019. DOI: 10.1007/s10115-018-1236-4.
- [30] B. Liu, “Sentiment analysis and opinion mining,” vol. 5, May 2012, ISBN: 978-3-642-19459-7. DOI: 10.2200/S00416ED1V01Y201204HLT016.
- [31] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques,” in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, Jul. 2002, pp. 79–86. DOI: 10.3115/1118693.1118704. [Online]. Available: <https://aclanthology.org/W02-1011>.
- [32] Statista, *Reddit users by country*, 2020. [Online]. Available: <https://www.statista.com/forecasts/1174696/reddit-user-by-country> (visited on 04/14/2021).
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: <https://aclanthology.org/N19-1423>.

- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010, ISBN: 9781510860964.
- [35] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter,” *CoRR*, vol. abs/1910.01108, 2019. arXiv: 1910.01108. [Online]. Available: <http://arxiv.org/abs/1910.01108>.
- [36] J. P. Ramos, R. J. Lopes, and D. Araújo, “Interactions between soccer teams reveal both design and emergence: Cooperation, competition and zipf-mandelbrot regularity,” *Chaos, Solitons & Fractals*, vol. 137, p. 109872, 2020, ISSN: 0960-0779. DOI: <https://doi.org/10.1016/j.chaos.2020.109872>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960077920302721>.