Developing a Portuguese Football Comprehensive Search System: A Multifaceted Approach

Alexandre Nunes
up202005358@edu.fe.up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

André Sousa
up202005277@edu.fe.up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

Gonçalo Pinto
up202004907@edu.fe.up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

Pedro Fonseca
up202008307@edu.fe.up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

ABSTRACT

Globally, an increase in data production is observed, posing the challenge of staying updated on topics of interest. To address this issue, a multifaceted search system is being developed, with a primary focus on the domain of Portuguese Football League news. In this initial milestone, the objective is to collect data from various sources and execute basic data processing. The data is sourced from platforms like Kaggle, Wikipedia, and digital newspapers such as Record, O Jogo, and A Bola, using web scraping techniques. Subsequently, this data is integrated into a centralized database, comprising 28,599 news articles, 3,414 game reports, and information on 33 football teams.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Web crawling**; **Document structure**.

KEYWORDS

Datasets, Data Preparation, Web Scraping, Web Crawling, Search System

ACM Reference Format:
Alexandre Nunes, André Sousa, Gonçalo Pinto, and Pedro Fonseca. 2023. Developing a Portuguese Football Comprehensive Search System: A Multifaceted Approach. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the digital age, the very fabric of our society has been reshaped by the proliferation of information. An era defined by the rapid evolution of technology and its profound influence on how information is accessed, consumed, and interacted with has been ushered in by what is commonly referred to as the "information society." These changes, particularly in sports, where the convergence of fans' passion, analysts' scrutiny, and researchers' curiosity[3] for timely and relevant sports news is leading to an increase in the size of the sports market, which is projected to reach $3.71 billion by 2027[12], have been particularly pronounced. In this dynamic landscape, a fundamental need within the realm of sports information is aimed to be addressed by the project. At its core, the development of a comprehensive sports news search system tailored specifically to the Portuguese League is sought. The development process of the previously described system is the objective of this paper. The inner workings of the data collection system will be explained first, followed by a thorough explanation of the pipeline and the subsequent work aimed to be developed. Finally, the datasets collected will be characterized, and the information needs will be enumerated.

2 DATA COLLECTION

The creation of a useful Portuguese football search system involved the development of a multifaceted data collection system. In this section, it is outlined the sources and methods employed to curate a diverse and extensive database that supports the search system's functionality.

2.1 Web Scraping from Record, O Jogo and A Bola

The arquivo.pt API[1] was the main source to gather information from different digital sports newspapers. This source allowed time travel on these websites which enriched the dataset with information from a longer range of time. Python[9] was used to build the pipeline that automatically connects to the arquivo.pt API to scrap Web pages of Record [10], O Jogo[6] and A Bola[2]. From these sources, news titles, bodies, and publication dates were extracted and then saved in a SQLite [11] database for later processing. This process resulted in 23593 documents created with information scrapped from Record, 1854 documents from O Jogo, and 3152 documents from A Bola.

2.2 Kaggle Dataset - Football Game Events

To complement the database with a granular perspective on football events within the Portuguese League, a meticulously curated dataset [5], made available in CSV format, encapsulating the main events of each football game spanning a decade was acquired from Kaggle [7].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

This dataset consists of data from 18 different football leagues (10 different countries), where each row of the dataset corresponds to data from a single game, containing multiple columns, but the most relevant ones are league name, teams' names, description of game events, game date, and final result. The original dataset had 96338 game instances, but since the focus was solely on the Portuguese League, the final dimension was 3414, providing information from all the matches between 2010 and 2022. Furthermore, the data was in pristine condition, as it did not contain missing values, and all the attributes were properly filled. Lastly, this information was made available under the Data files © Original Authors license, with the usage for commercial purposes being forbidden. With this dataset, the system can once again be enriched to provide information about specific in-game events that may not have been covered by the news articles that were collected. By integrating this data, the aim was to enhance the depth and comprehensiveness of the search engine's offerings.

2.3 Wikipedia - Football Team Histories

Since the goal is not to be limited to sports news only, Wikipedia [13] was used to enrich the search system with rich textual data that tell the history of all the teams of the Portuguese League. By storing this data in the previously mentioned database, a greater range of information needs could be met. A total of 33 different Portuguese teams, representing all the teams featured in the previously mentioned dataset, were gathered. Collectively, the data collection process involved a blend of manual curation, web scraping, and dataset acquisition from reputable sources. The convergence of these diverse data streams forms the foundations of the search system, ensuring that a rich and multifaceted panorama of sports news and historical insights related to the Portuguese League is encapsulated. The subsequent sections of this article will focus on the preprocessing and organization of this data to realize the search system's functionality and utility within the information society.

2.4 Pipeline description

As shown in **Figure 1**, the data collection pipeline has three main starting points that together build the foundations of the search system. One of these starting points corresponds to the aforementioned Kaggle dataset. This dataset was previously downloaded and processed by the pipeline to build rich textual descriptions of each football game (including goals, faults, teams, and player identifications) using pre-defined sentences to help make the dataset's information more appealing to the end user. The final result of this stage is stored in the document collection called **game_report**. The pipeline also accesses arquivo.pt to retrieve trustworthy news articles from websites like Record, O Jogo, and A Bola. This process required some fine-tuning due to the limitations of the arquivo.pt API: there is a limit of 250 requests per minute, therefore a custom class was created, thus enabling the API requests to be made without exceeding the limit. Following that, the articles are scrapped using BeautifulSoup [4] to create documents containing only the relevant information, which includes the title, body, and publishing date (due to different date formats used by different websites, normalization was required). An experimental feature of this project is the capability to process not only textual news but also video content. This involves downloading the videos to local storage, uploading them to the OpenAI's API that performs speech-to-text conversion (Whisper [8]), and ultimately adding these news items to the textual ones. These documents were stored in the **article** collection. Finally, to further complement the data collection, the pipeline also includes data retrieved from the history section of the Wikipedia page of each football team, creating a new collection of documents called **team_info**. The last step of the pipeline is storing the collections in the SQLite database.

To simplify the execution process, a **Makefile** is available to run the entire pipeline. It offers two execution modes: **run**, which processes the complete pipeline, and **partial**, designed for a simplified demonstration of the pipeline. Additionally, a **help** command is provided for improved documentation, as well as a **install** command used to install all the required Python packages.

2.5 Information needs

As with any search system, the goal is to answer the queries of the target audience. These queries may vary from highly specific requests, such as the outcome of a football match, to more open-ended and ambiguous topics related to football, like "Biggest comeback in Portuguese league". In this context, users might input various search scenarios like:

- SC Braga coach controversy.
- Biggest transfer of 2015.
- Poor refereeing performance in important matches.
- Result of the 2015 Porto vs. Benfica match.
- Jonas's goal in minute 97.

2.6 Conceptual Data Model

By the illustration in **Figure 2**, the foundation of the search system relies on a database of textual documents that aggregate the information needed to answer the user's queries. This information covers different entities and relationships. At its core, news articles are gathered, each associated with a given website (the one where the article was published). Each news article is characterized by its title, body, date of publication, and the publisher's name. It is worth noting that a news article may be associated with several teams and also several players. Moreover, the data stored about the teams contains their names and histories, as well as their associations with the players that are part of each team. Finally, the system also gathers the match's final result, the name of the team that is playing at home and the name of the one that isn't, and important events such as goals, faults, etc., that are associated with the two teams at a time. Lastly, short texts are generated using this information to provide the user with descriptions of the main events of the game.

2.7 Dataset Characterization

The dataset under scrutiny, comprising approximately 30,000 news articles meticulously aggregated from esteemed Portuguese media outlets such as Record, O Jogo, and A Bola, provides a comprehensive panorama of the Portuguese football landscape. This extensive collection spans a substantial timeframe, offering valuable insights into the evolution of media coverage from 2007 to 2022.

Upon reviewing the Data Summary plot (**Figure 3**), which succinctly encapsulates the dataset's essentials, it becomes evident that the dataset not only encompasses news articles but also incorporates Wikipedia descriptions, game reports, and even the average word count per text. This diverse range of content types enriches the dataset, offering multifaceted perspectives on Portuguese football.

A meticulous analysis of the dataset's temporal distribution, as showcased in the Time Graph (**Figure 4**), reveals intriguing patterns. Particularly noteworthy is the substantial surge in collected news articles observed from 2017 to 2020, with a distinct spike in 2010. These temporal nuances illuminate periods of heightened media activity, potentially aligning with significant football events or developments within the Portuguese League. Another possible reason is the fact that arquivo.pt does not store all the Web pages consistently, therefore it is normal to have some years with discrepancies in the number of pages.

Looking into the dataset's content, the prominence of specific football entities becomes apparent. Major football clubs such as Benfica, Sporting, and FC Porto dominate the dataset's mentions, as highlighted in the Popular Teams plot (**Figure 5**). This plot provides a clear visualization of the teams that command significant media attention, offering valuable insights into the media's focal points.

Furthermore, the Top 20 Popular Entities plot (**Figure 6**) presents a comprehensive list of the most frequently mentioned entities in the news articles. This detailed breakdown, obtained using Natural Language Processing Entity Extraction techniques, sheds light on the diverse array of entities that constitute the dataset, revealing the depth of topics covered within the Portuguese football domain.

A nuanced exploration of the textual content in the dataset is facilitated by the WordCloud visualization (**Figure 7**), which encapsulates the most frequently mentioned words. This word cloud offers a visual representation of the dataset's thematic undercurrents, highlighting the key topics and terminologies that permeate the Portuguese football narratives.

Adding a layer of geographical context, the dataset incorporates pertinent information about the regions associated with Portuguese football teams. This contextual information, sourced from Wikipedia descriptions, enriches the dataset significantly, providing valuable insights into the regional affiliations and historical backgrounds of the football clubs featured in the news articles. Such geographical context augments the dataset's comprehensiveness, offering users a holistic understanding of the diverse landscapes from which these football clubs emerge.

To summarize, this meticulously curated dataset offers a comprehensive view of the dynamic evolution of Portuguese football. Through extensive validation, detailed textual analysis, and the incorporation of geographical context, the dataset emerges as a robust and multifaceted resource. It serves as the foundation for the development of a comprehensive sports news search system, empowering users to explore the intricate world of Portuguese football with precision and depth.

2.8 Document Representation

The developed search system will answer each user query with a list of documents. The main structure of a document consists of its title, contents, and publishing date. In the case of documents generated from match reports, the title will be the concatenation of the names of the two teams involved, as well as its final result, the content will be the report of the game and, finally, the publishing date will be the date of the match. Lastly, the documents generated with the Wikipedia data will be used as a complement to the remaining documents, to enrich the aforementioned documents.

2.9 Conclusions

In this paper, the foundation for a comprehensive sports news search system dedicated to the Portuguese Football League has been laid. A rich and multifaceted dataset was successfully gathered from diverse sources, creating a resource that serves the information needs of sports enthusiasts. The data collection, representation, and analysis have provided valuable insights into the dynamic world of Portuguese football, representing the source material of the upcoming search system. As the project continues to develop, the aim is to empower users with the tools to explore this domain with depth and precision.

REFERENCES

[1] Arquivo.pt. Arquivo.pt homepage. https://arquivo.pt/, 2023. [Online; acedido 1 de outubro de 2023].

[2] A Bola. A bola homepage. https://www.abola.pt/, 2023. [Online; acedido 1 de outubro de 2023].

[3] Preethi Cheguri. Data science in football: The world cup is more data-focused now. https://www.analyticsinsight.net/data-science-in-football-the-world-cup-is-more-data-focused-now/, 2023. [Online; acedido 1 de outubro de 2023].

[4] Crummy. Beautiful soup documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/, 2023. [Online; acedido 1 de outubro de 2023].

[5] Sebastian Gebala. Football dataset +96k matches (18 leagues). https://www.kaggle.com/datasets/bastekforever/complete-football-data-89000-matches-18-leagues, 2023. [Online; acedido 1 de outubro de 2023].

[6] O Jogo. O jogo homepage. https://www.ojogo.pt/, 2023. [Online; acedido 1 de outubro de 2023].

[7] Kaggle. Kaggle homepage. https://www.kaggle.com/, 2023. [Online; acedido 1 de outubro de 2023].

[8] OpenAI. Introducing whisper. https://openai.com/research/whisper, 2023. [Online; acedido 1 de outubro de 2023].

[9] Python. Python homepage. https://www.python.org/, 2023. [Online; acedido 1 de outubro de 2023].

[10] Record. Record homepage. https://www.record.pt/, 2023. [Online; acedido 1 de outubro de 2023].

[11] SQLite. Sqlite homepage. https://www.sqlite.org/index.html, 2023. [Online; acedido 1 de outubro de 2023].

[12] Statista. Sports - worldwide. https://www.statista.com/outlook/dmo/app/sports/worldwide?currency=USD, 2023. [Online; acedido 1 de outubro de 2023].

[13] Wikipedia. Wikipedia welcome page. https://pt.wikipedia.org/, 2023. [Online; acedido 1 de outubro de 2023].

A FIGURES

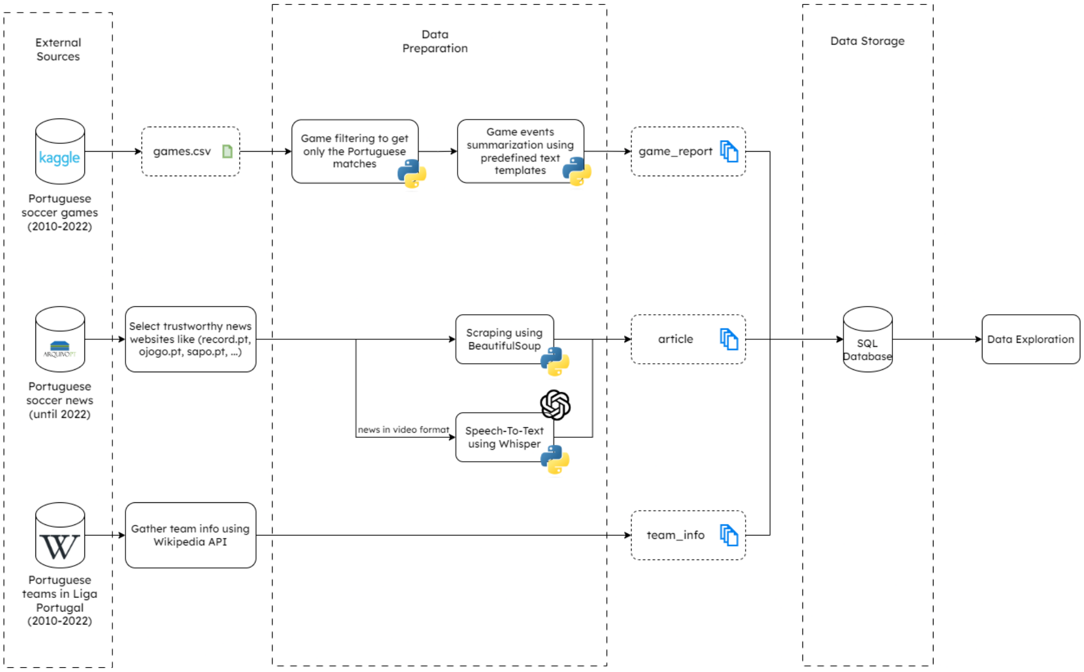
External Sources
kaggle
Portuguese soccer games (2010-2022)
ARQUIVO.PT
Portuguese soccer news (until 2022)
W
Portuguese teams in Liga Portugal (2010-2022)
games.csv
Select trustworthy news websites like (record.pt, ojogo.pt, sapo.pt, ...)
Gather team info using Wikipedia API
Data Preparation
Game filtering to get only the Portuguese matches
Game events summarization using predefined text templates
Scraping using BeautifulSoup
news in video format
Speech-To-Text using Whisper
game_report
article
team_info
Data Storage
SQL Database
Data Exploration

Figure 1: Data Pipeline

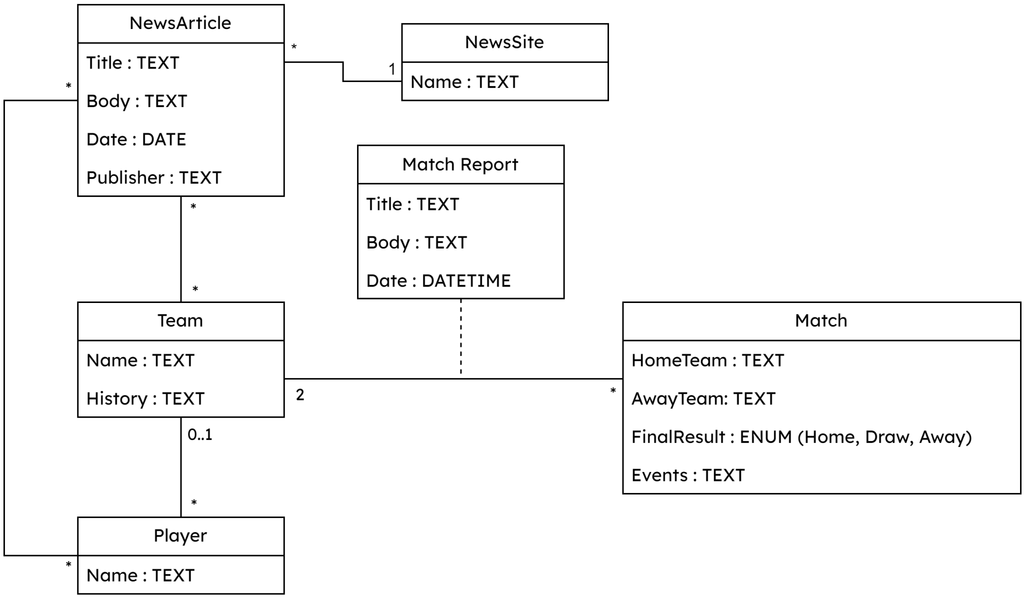

Figure 2: Conceptual Data Model

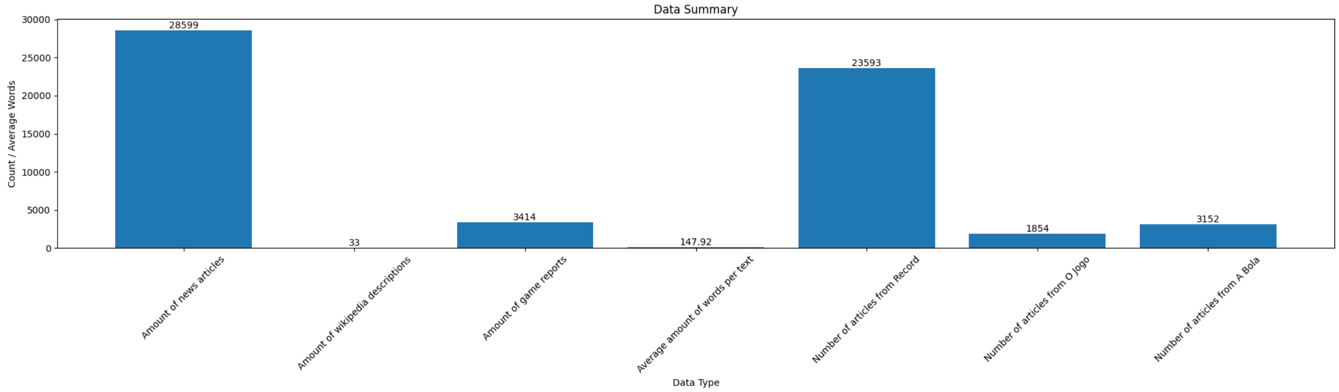

Figure 3: Data Summary

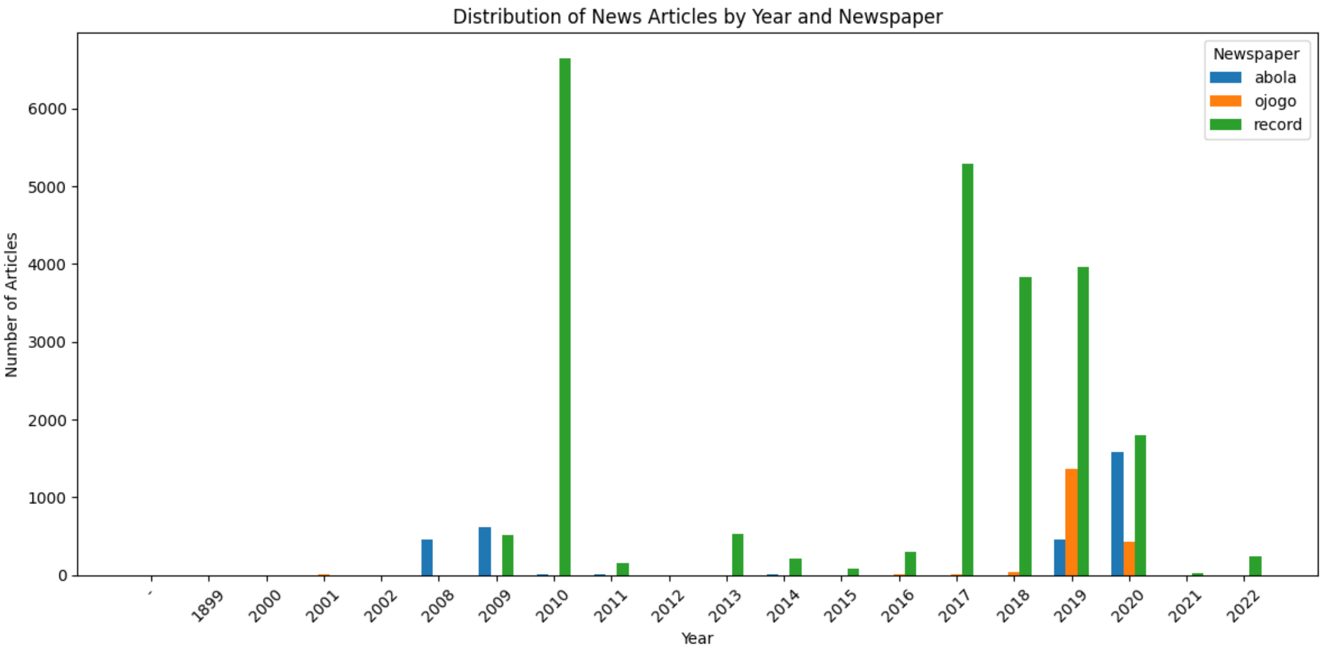

Figure 4: Time Graph

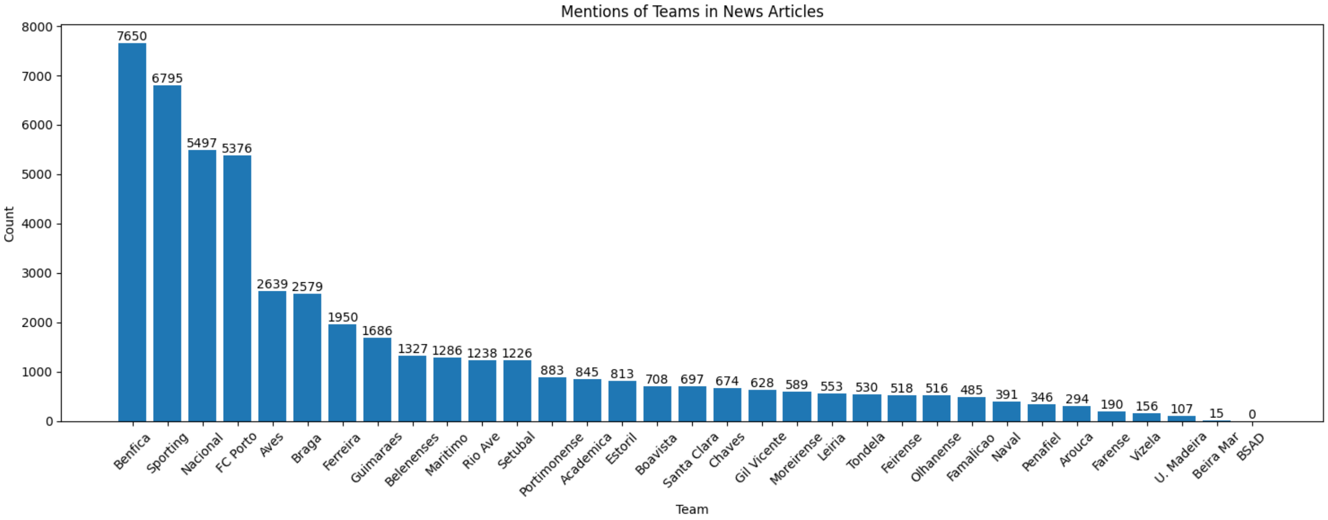

Figure 5: Popular Teams

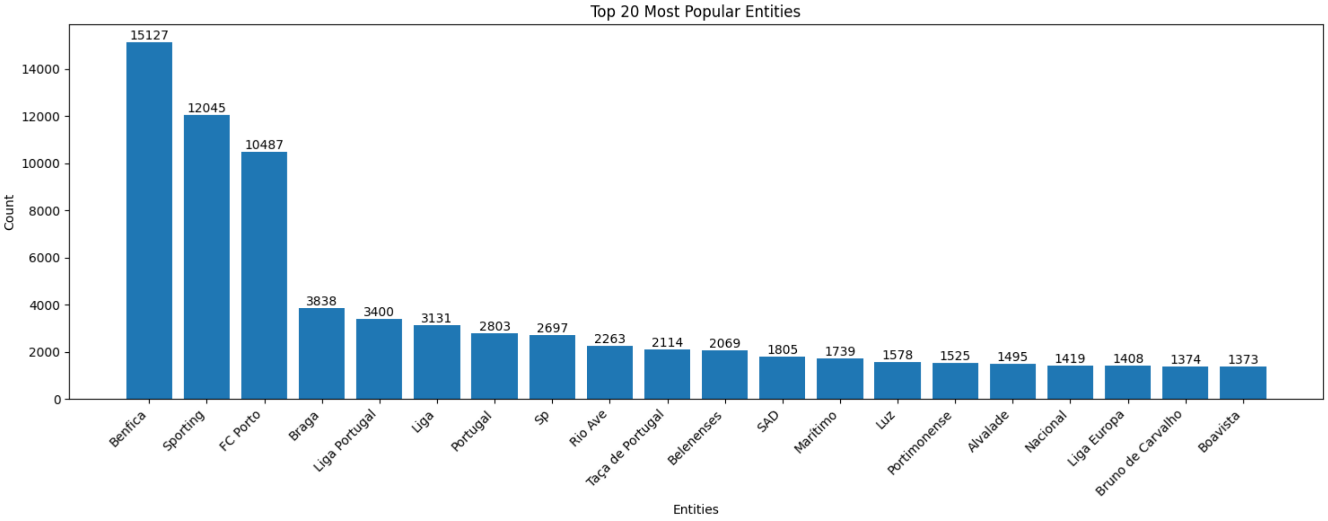

Figure 6: Top 20 Most Popular Entities

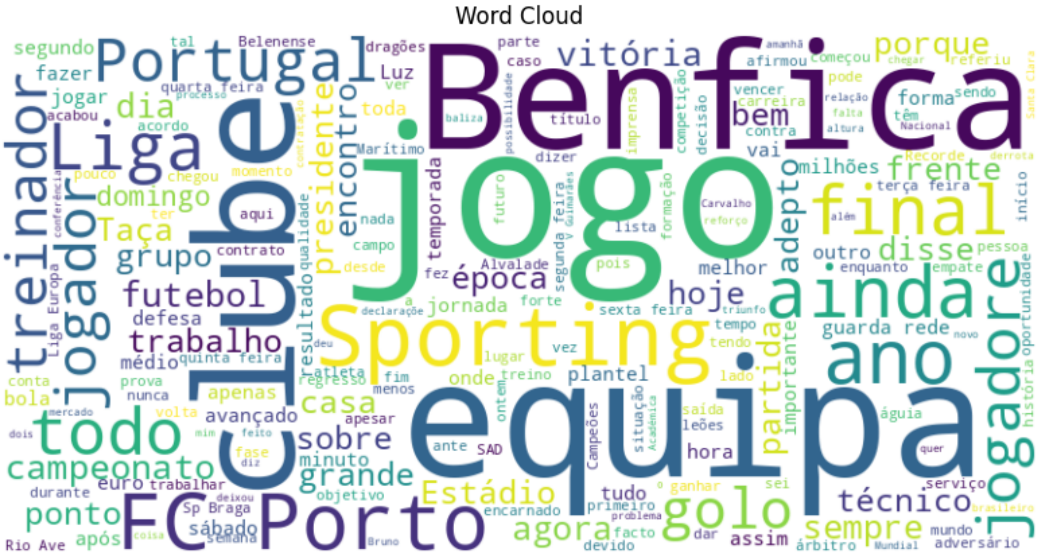

Figure 7: WordCloud