Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Processamento de Linguagem Natural aplicado a notícias futebolísticas

Murilo Rejani Franzotti

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Processamento de Linguagem Natural aplicado a notícias futebolísticas

Murilo Rejani Franzotti
Orientador(a): Thiago Rodrigo Ramos

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Natural Language Processing applied to football news

Murilo Rejani Franzotti
Advisor: Thiago Rodrigo Ramos

Bachelors dissertation submitted to the Depart ment of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
November of 2025

Murilo Rejani Franzotti

Processamento de Linguagem Natural aplicado a notícias futebolísticas

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Murilo Rejani Franzotti e aprovado pela banca examinadora.

Aprovado em 28 de Novembro de 2025

Banca Examinadora:

- Thiago Rodrigo Ramos
- Rafael Izbicki
- Márcio Alves Diniz

A todos que contribuíram para a minha formação.

Agradecimentos

A conclusão desta fase da minha vida marca o início da minha carreira profissional. Aqui se encerra um período marcado por novas experiências, estudantis e particulares, que enriqueceram e proporcionaram a mim novos e valiosos aprendizados. Na faculdade, nos deparamos com um mundo totalmente novo, ainda mais para mim, mudando de cidade e de convívio social. Isso, de fato, se tornou o desbravamento de novos ares e lugares.

A minha trajetória foi marcada de conquistas e feitos pessoais, que com esforço e dedicação alcancei durante a graduação. Não poderia deixar de agradecer a todos com quem tive convívio durante este tempo e a Deus, pela força e saúde que Ele me deu. Aos familiares, que a todo tempo me serviram de base para que não desmoronasse. Aos amigos, que me proporcionaram momentos incríveis, desde os estudos até os de lazer. Aos professores, por me ajudarem a conhecer este novo mundo que é a Estatística.

A graduação é, para todo estudante, um período de transformação. Se conheça, mas não saia da sua realidade e da sua personalidade. A você, estudante, que está lendo, aproveite cada momento da faculdade, eles são únicos. A você, formado, lembre-se dos momentos que se passaram na universidade, tenho certeza que você os guarda com alegria no seu coração.

Sem dados, você é apenas mais uma pessoa com uma opinião.

(W. Edwards Deming)

Resumo

O Processamento de Linguagem Natural (PLN) é um importante conjunto de técnicas e ferramentas aptas para as análises de textos e somado ao uso eficiente dos procedimentos de coleta de dados da internet, *Web Scraping*, formam processos eficazes para as análises deles. No contexto do futebol, as notícias esportivas exercem grande relevância nos acontecimentos, pois descrevem e, muitas vezes, causam tais fatos. Analisar estes textos resulta em indicativos significativos sobre os causos do futebol, pois refletem as vontades e os pensamentos dos torcedores. Além disso, temporalmente conseguimos distinguir quando um fato impacta o mundo do futebol, seja pelo aumento do número de notícias, seja pelo aumento das relações de uma palavra com a outra. Com isso em mente, analisar a similaridade entre jogadores e técnicos com seus respectivos clubes e seleções traz conclusões sobre o quão impactante foi o evento para o público que acompanha o esporte.

Logo, com as notícias conseguimos ter bons indicativos entre o final do ano de 2023 até o penúltimo trimestre de 2025. Vemos, por exemplo, que o fato do jogador Neymar ter se transferido para o Santos corroborou para o seu aporte midiático e com o Messi indo atuar nos Estados Unidos aumenta-se os rumores de sua aposentadoria, pois os holofotes não estão mais sobre o jogador. Além disso, títulos aumentam significativamente a aparição de um clube na mídia, como as conquistas do Campeonato Brasileiro e da Copa Libertadores pelo Botafogo, no final de 2024. Estes e outros acontecimentos impactam o mundo do futebol e conseguimos analisar isto com este tipo de dado, as notícias.

Palavras-chave: *Web Scraping, Processamento de Linguagem Natural, Notícias, Futebol.*

Abstract

Natural Language Processing (NLP) is an important set of techniques and tools suitable for text analysis and, when combined with the efficient use of data collection procedures from the internet, web scraping, they form effective processes for their analysis. In the context of football, sports news plays a significant role in events, as it describes and, in many cases, causes such facts. Analyzing these texts results in significant indicators about football-related events, as they reflect the desires and thoughts of fans. Furthermore, from a temporal perspective, it is possible to distinguish when an event impacts the football world, either through an increase in the number of news articles or through the strengthening of relationships between one word and another. With this in mind, analyzing the similarity between players and coaches and their respective clubs and national teams leads to conclusions regarding how impactful an event was for the audience that follows the sport.

Thus, through news articles, it is possible to obtain relevant indicators from the end of 2023 until the penultimate quarter of 2025. For example, the fact that the player Neymar transferred to Santos corroborated his media exposure, and with Messi moving to play in the United States, rumors of his retirement increased, as the spotlight is no longer focused on the player. In addition, titles significantly increase a club's appearance in the media, such as Botafogo's victories in the Brazilian Championship and the Copa Libertadores at the end of 2024. These and other events impact the football world, and it is possible to analyze this through this type of data, namely, news articles.

Keywords: *Web Scraping, Natural Language Process, News, Football.*

Lista de Figuras

2.1 Fluxograma das etapas de coleta e transformação de metadados. 30
2.2 Exemplo de um *lobby* da plataforma 90min (2025). 32
2.3 Notícia de exemplo na página digital. 38

3.1 Processo da função *Word2Vec* para vetorização. 45
3.2 Exemplo da operação matemática entre palavras. 47
3.3 Exemplo de similaridades entre as palavras. 48

4.1 Número de notícias por trimestre. 54
4.2 Entidades mais citadas. 55
4.3 Número de notícias com a palavra "Neymar". 55
4.4 Número de notícias com a palavra "Lionel Messi". 56
4.5 Número de notícias com a palavra "Liverpool". 57
4.6 Número de notícias com a palavra "Botafogo". 57
4.7 Top 10 palavras mais similares a "Neymar". 58
4.8 Similaridade de "Neymar"com clubes. 59
4.9 Similaridade de "Neymar"com termos que remetem à lesão. 60
4.10 Top 10 palavras mais similares a "Messi". 61
4.11 Similaridade de "Messi"com clubes. 61
4.12 Top 10 palavras mais similares a "Liverpool". 62
4.13 Similaridade de "Liverpool"com jogadores. 63
4.14 Top 10 palavras mais similares a "Botafogo". 64
4.15 Similaridade de "Botafogo"com campeonatos. 65
4.16 PCA para entidades PER. 66
4.17 PCA para entidades LOC. 66
4.18 PCA para entidades ORG. 67
4.19 Clusterização do t-SNE para entidades PER. 68

4.20 Clusterização do t-SNE para entidades LOC. 69
4.21 Clusterização do t-SNE para entidades ORG. 70
4.22 Equação para a palavra "Mbappe". 71
4.23 Equação para a palavra "Dorival". 72
4.24 Equação para a palavra "Mirassol". 72

Lista de Tabelas

2.1 Exemplo de observações no banco de dados final. 39

Lista de Códigos

2.1 Código HTML do *lobby* da página 160 do site 90min (2025). 32
2.2 Captura do local das URL's do *lobby* da página 160. 34
2.3 Captura das URL's do 160º *lobby*. 35
2.4 Exemplo de uma notícia no *lobby* em formato JSON. 35
2.5 Coleta dos metadados das notícias. 36
2.6 Criação do Dataset de notícias. 37
2.7 Notícia de exemplo em JSON. 38
3.1 Texto antes do pré-processamento. 42
3.2 Texto após do pré-processamento. 43
3.3 Aplicação do NER nos textos. 49

Sumário

1 Introdução **25**

1.1 Objetivo 27

1.2 Organização do Trabalho 27

2 Coleta de Dados **29**

2.1 *Web Scraping* 29

2.2 Processo de extração dos dados 32

2.2.1 Acesso ao *lobby* 32

2.2.2 Coletando os links 34

2.2.3 Acesso à notícia 36

2.2.4 Banco de Dados final 37

3 Processamento **41**

3.1 Pré-processamento 41

3.2 Vetorização 43

3.3 Análise de Similaridade 47

3.4 Identificação de Entidades 49

3.5 Redução de Dimensionalidade 50

3.5.1 Análise de Componentes Principais 51

3.5.2 t-SNE 52

4 Resultados **53**

4.1 Análise Descritiva 53

4.2 Análises das Similaridades 58

4.3 Reduções de Dimensionalidade 65

4.4 Equações matemáticas 70

5 Conclusão **75**

Referências Bibliográficas **75**

Capítulo 1

Introdução

A Estatística é uma área das ciências exatas que permeia diversos cenários, desempenhando um papel fundamental na análise e na compreensão de fenômenos a partir das informações que os caracterizam. No cotidiano, os desafios enfrentados por um estatístico podem ser complexos, seja pela escassez de dados, pela natureza intrínseca deles ou por limitações computacionais. Embora grande parte dessas limitações tenham sido superadas ao longo deste século, graças ao avanço no processamento e no armazenamento de informações, certas categorias de dados permaneceram praticamente inacessíveis no passado, seja pela dificuldade de coleta, seja pelo alto custo computacional de sua análise.

Entre essas categorias, o processamento e a análise de palavras têm despertado interesse desde meados do século XX. As primeiras pesquisas sistemáticas surgiram na década de 1950: Hutchins (2004) descreve o Experimento de Georgetown, conduzido pela IBM, que apresentou um dos primeiros algoritmos de tradução automática. A partir daí, novos estudos foram sendo publicados e técnicas gradualmente aprimoradas. Mais tarde, Martin e Jurafsky (2025) consolidaram esse campo ao formalizá-lo como Processamento de Linguagem Natural (PLN) em *Speech and Language Processing*. Com o surgimento e popularização das redes sociais, a aplicabilidade do PLN se tornou ainda mais ampla, facilitando o acesso a dados textuais em larga escala e impulsionando a investigação de novos problemas e aplicações, como ilustrado em estudos recentes (da Silva Ferreira e Sampaio (2018)).

Neste contexto de crescente disponibilidade de dados textuais, o avanço das técnicas de Aprendizado de Máquina (AM) desempenhou um papel determinante. Essa evolução tornou possível não apenas tarefas preditivas, mas também análises mais profundas sobre diferentes línguas, estruturas frasais e textos provenientes dos mais variados domínios.

Izbicki e dos Santos (2022) discutem as principais técnicas de AM que, à primeira vista, parecem aplicáveis apenas a dados numéricos; contudo, ao longo desta monografia, veremos como tais técnicas podem ser adaptadas para inúmeras categorias de informação, bastando para isso transformações adequadas.

O interesse em compreender automaticamente textos não é recente. Schank e Abelson (1977) introduziram a ideia de compreensão automática da linguagem, que posteriormente inspiraria técnicas de monitoramento, interpretação e identificação de tendências em conteúdos relevantes. No cenário contemporâneo, marcado pelo fluxo constante de informações em portais jornalísticos, o PLN tornou-se ainda mais relevante, permitindo análises sobre muitos aspectos, como a credibilidade textual (Sudhakar e Kaliyamurthie (2024)).

De fato, textos jornalísticos moldam percepções e influenciam opiniões, muitas vezes por meio de escolhas sutis de vocabulário. Martin e Jurafsky (2025) mostram, no capítulo 22, como a substituição de palavras semanticamente semelhantes, mas com graus distintos de intensidade, pode direcionar a interpretação do leitor. Essa estratégia de persuasão está presente em diferentes segmentos midiáticos. No futebol, por exemplo, rumores de conflitos entre jogadores e técnicos, noticiados meses antes, muitas vezes antecedem demissões, afastamentos ou especulações sobre transferências, cuja força depende diretamente da reação conjunta da imprensa e dos torcedores.

Com o advento das redes sociais, esse fenômeno ampliou-se significativamente. A interação direta entre jogadores, jornalistas e torcedores gerou um aumento expressivo de dados textuais, possibilitando estudos que demonstram que tais textos não apenas comunicam fatos, mas também refletem tendências, percepções coletivas e humores sociais. Afonso e Duque (2020), por exemplo, investigam como as frequências de determinadas palavras variaram durante os primeiros meses da pandemia de Covid-19, em resposta aos acontecimentos daquele período. De forma análoga, no contexto futebolístico, torna-se possível analisar o comportamento de termos específicos e suas relações, uma vertente ainda pouco explorada na literatura, mas que vem ganhando relevância com o aumento do público e das novas dinâmicas do mercado esportivo.

Em síntese, todas estas manifestações textuais configuram um ecossistema rico que permite investigar padrões, mudanças de significado e comportamentos linguísticos associados a determinados temas. Assim, este trabalho buscará identificar e analisar tais relações no ambiente futebolístico, investigando, por exemplo, a influência de determi-

nados profissionais dentro de seus clubes, o crescimento ou declínio do desempenho de equipes no cenário nacional e internacional, bem como possíveis alterações no sentido das palavras empregadas em textos de cunho esportivo. Os códigos de todas as análises que serão relatadas neste trabalho podem ser encontradas na plataforma *GitHub* através do link https://github.com/Murilo003/Trabalho-Graduacao.

1.1 Objetivo

O objetivo desta monografia é discorrer acerca das técnicas de *Processamento de Linguagem Natural* e aplicá-las dentro do contexto futebolístico, com o intuito de buscar padrões linguísticos e relações entre os termos ou expressões. Para isso, as técnicas que serão utilizadas levarão em consideração, principalmente, as relações que as palavras tem entre si, fazendo uso delas em uma visão geral e discorrendo pontualmente acerca de alguns termos que forem julgados ideias para a análise.

1.2 Organização do Trabalho

A monografia está estruturada da seguinte forma:

No Capítulo 2, é apresentado o método de coleta dos dados, bem como o portal de notícias esportivas utilizado como fonte das informações e a composição do conjunto de dados coletado. Em seguida, o Capítulo 3 descreve as técnicas de Processamento de Linguagem Natural empregadas ao longo do trabalho, tais como a remoção de stopwords e a vetorização de palavras, além dos conceitos estatísticos que fundamentam as análises textuais realizadas. Posteriormente, no Capítulo 4, são apresentados e discutidos os resultados obtidos a partir da aplicação dessas técnicas às palavras consideradas pertinentes ao tema em estudo. Por fim, o Capítulo 5 conclui sobre as técnicas abordadas no estudo e sintetiza os resultados encontrados no trabalho.

Capítulo 2

Coleta de Dados

Neste capítulo, descreveremos sobre os dados utilizados neste estudo. Coletados pelo autor, as informações são oriundas do portal de notícias esportivas 90min (2025), composto por diversos assuntos relacionados ao futebol brasileiro e internacional, de textos informando sobre transferências de jogadores até os resultados dos jogos de diversas competições. O período adotado para a coleta compreende o último trimestre de 2023 até o terceiro trimestre de 2025. Essa janela foi escolhida por apresentar um número de notícias adequado de publicações para as análises a serem feitas, além de abranger eventos relevantes como o encerramento de temporadas, janelas de transferências nacionais e internacionais e torneios continentais. Tais características garantem diversidade temática e temporal nos dados, o que contribui para análises mais amplas sobre o comportamento linguístico.

2.1 *Web Scraping*

As técnicas de coleta de dados na *internet* só foram possíveis após a década de 1990. Com a criação do *World Wide Web*, ou "www", as plataformas de pesquisa foram formadas para dar recomendações de outros sites através do endereço digital. E, para que essas recomendações fossem assertivas, tais plataformas precisavam acessar as páginas digitais e coletar seu conteúdo. Kosala e Blockeel (2000) introduzem as técnias de *Web Mining*, que mais tarde seriam aprimoradas para as técnicas de *Web Scraping*.

Com os avanços da programação, estas técnicas se tornaram palpáveis aos usuários das mais diversas linguagens. Em Python, a biblioteca *requests* (Cordasco *et al.* (2024)) permite que essa coleta seja feita de maneira eficaz e é amplamente utilizada para requisi-

ções HTTP, que resultam em todas as informações contidas na página digital, e extração de conteúdo da *web* de forma automatizada.

No geral, a identificação da linguagem em que o site foi feito rege os passos do *Web Scraping*, mas esse processo tende a seguir por uma sequência de etapas fundamentais. A primeira delas é a identificação do site-alvo, ou seja, a definição da página da *web* da qual se deseja extrair os dados. Em seguida, realiza-se a seleção das informações relevantes, priorizando os metadados que serão extraídos conforme os objetivos da análise. A etapa seguinte é a de extração, na qual, com o auxílio de softwares, os dados são coletados diretamente do código do site, considerando tanto a linguagem utilizada na construção da página quanto a estrutura onde essas informações estão localizadas. Por fim, ocorre a transformação dos dados brutos para formatos apropriados ao processamento e análise, sendo os formatos JSON e CSV os comumente utilizados. Este procedimento pode ser visualizado no fluxograma da Figura 2.1.

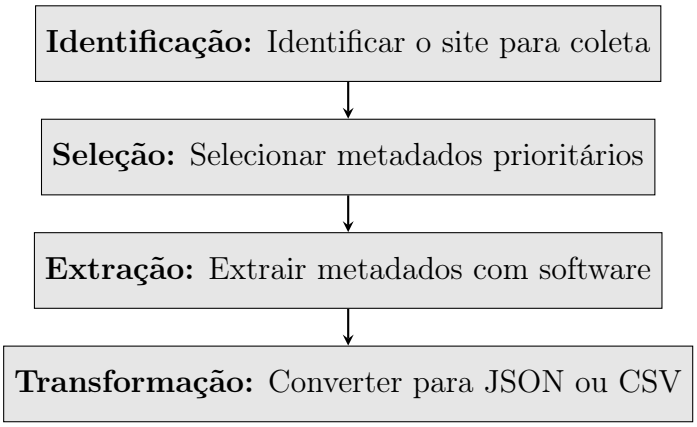

Figura 2.1: Fluxograma das etapas de coleta e transformação de metadados.

No contexto do PLN, o foco principal está na análise textual. Por isso, identificar corretamente a estrutura dos parágrafos no código desenvolvido para as páginas digitais é o que permite a implementação destas técnicas. Uma vez compreendida, a coleta é feita de maneira eficaz, resultando em uma amostra idêntica ao visto na *web*. É importante ressaltar que conseguimos coletar as diversas informações que compõem uma página digital, não se limitando a textos. Em suma, tudo que está nas plataformas pode ser reunido para análise.

Portanto, a aplicação desta técnica para a coleta dos dados seguiu a estrutura supracitada e, de maneira similar, vamos elencar como foi discorrida esta etapa do processo.

- **Identificação:** há diversas plataformas digitais que propagam as notícias em es-

tudo. Logo existem também muitas maneiras de construção das mesmas. Escolher uma plataforma que facilite a coleta dos dados foi o principal critério usado neste trabalho. Páginas digitais desenvolvidas em JavaScript, por exemplo, são complexas por dificultar sua localização; já em HTML, a presença das *tags* (elementos que estruturam o conteúdo exibido) auxiliam na procura dos dados. Além disso, a opção de plataformas que possuem o carregamento de novas páginas por meio de URL's (*Uniform Resource Locator*) estruturadas sequencialmente (como ".../pagina/1" e ".../pagina/2") foi priorizado por sua simplicidade computacional na coleta dos dados em comparação as plataformas construídas com carregamento dinâmico do conteúdo, em geral, por rolagem infinita;

- **Seleção:** diferentemente das plataformas de outros ramos, as que divulgam notícias possuem um número de metadados disponíveis limitados. Portanto, foram priorizados o texto principal, o título, o autor, a data e o endereço da notícia;

  - **Título:** frase relativa ao título dado a notícia;
  - **Texto:** corpo textual da notícia;
  - **Link:** URL da notícia;
  - **Data:** data em que a notícia foi publicada na plataforma;
  - **Autor:** jornalista responsável pela escrita e publicação da notícia.

- **Extração:** a biblioteca *requests* foi usada para a extração. A plataforma selecionada é construída na linguagem HTML e com URL's sequenciais. Logo, houve uma maior facilidade computacional para retirar os endereços digitais de cada notícia e, em seguida, acessar separadamente cada notícia para a extração dos metadados selecionados, que serão salvos a priori na mesma linguagem do site;

- **Transformação:** por fim, os metadados serão transformados no formato JSON, um modelo de dicionário adequado para o armazenamento dos metadados, para em seguida serem armazenados em CSV, facilitando a condução dos dados entre os códigos utilizados na análise.

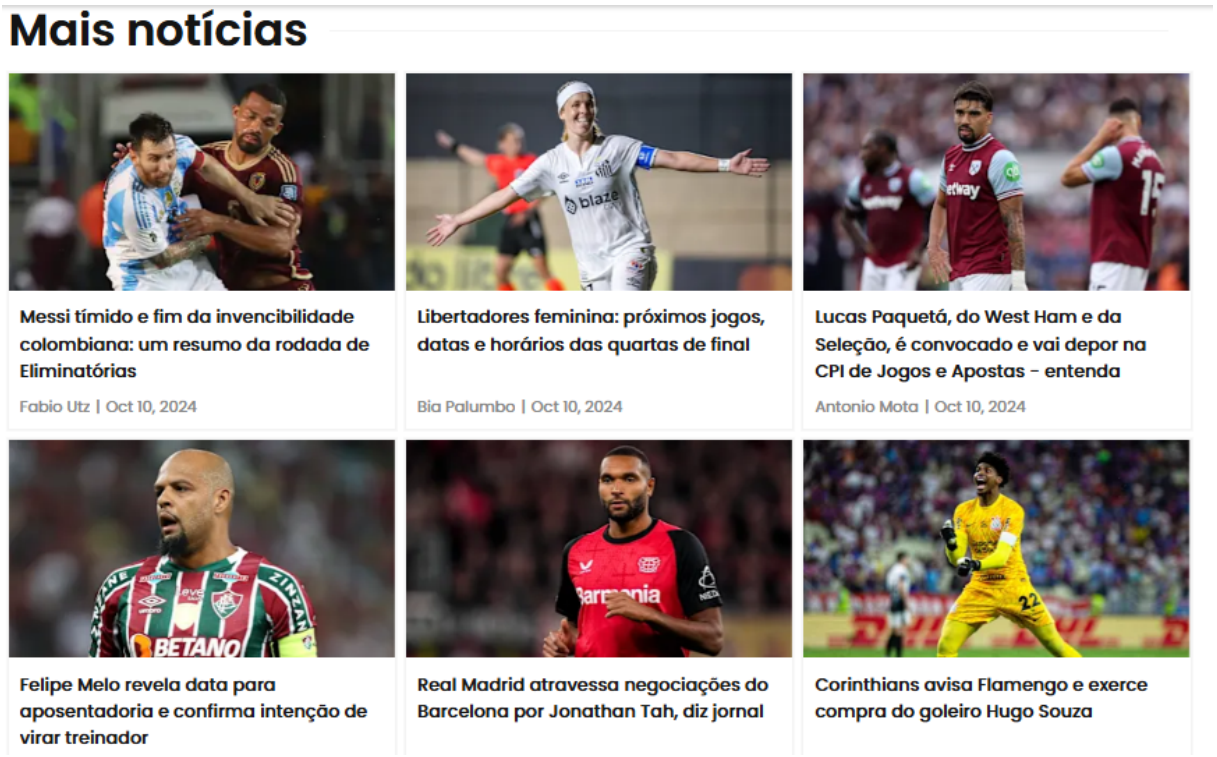

Figura 2.2: Exemplo de um *lobby* da plataforma 90min (2025).

2.2 Processo de extração dos dados

Os dados para este estudo são textos e outros metadados das notícias futebolísticas presentes na plataforma digital 90min (2025). É importante citar que plaformas de notícias como esta possuem uma estrutura particular, onde o primeiro acesso é a um *lobby*, local em que são apresentadas diversas notícias para o usuário selecionar a de seu interesse, como representado na Figura 2.2, e em seguida o acesso ao corpo da notícia de interesse do usuário. Observamos que os *lobbies* são ordenados, onde no primeiro se encontram as notícias mais recentes e, portanto, quanto maior for a ordem, mais antigas serão as notícias contidas nele. Tal acesso inicial exige que o processo de coleta possua uma etapa a mais.

2.2.1 Acesso ao *lobby*

Ao entrar na plataforma devemos ter em mente que o algoritmo não irá ler uma notícia como um leitor, mas irá acessar o código fonte da notícia e, a partir dela, coletar os textos; o primeiro acesso será ao código fonte em HTML do *lobby*.

```
1 <!DOCTYPE html>
2     <html  lang="pt-BR" kasda>
```

```
3     <head>
4         <meta charset="UTF-8">
5         <meta name="viewport" content="width=device-width, initial-
              scale=1">
6
7     <title>Destaques no 90min: tudo que de mais interessante rolou
          no futebol Page 160</title>
8     (...)
9     <script type="text/javascript"> (...)
10    <script type="application/ld+json"> (...){"@type":"NewsArticle"
          ,"position":9,"url":"https://www.90min.com/pt-br/corinthians
          -avisa-flamengo-prepara-compra-goleiro-hugo-souza","headline
          ":"Corinthians avisa Flamengo e exerce compra do goleiro
          Hugo Souza","image":"https://images2.minutemediacdn.com/
          image/upload/c_crop,w_4176,h_2349,x_0,y_269/c_fill,w_720,
          ar_16:9,f_auto,q_auto,g_auto/images/GettyImages/mmsport/90
          min_pt-BR_international_web/01j9vkdpb5k58m353qp0.jpg","
          datePublished":"2024-10-10T18:01:07Z","author":{"@type":"
          Person","name":"Fabrício Carvalho","url":"https://www.90min.
          com/pt-BR/authors/fabricio-ca"},"publisher":{"@type":"
          Organization","name":"90min.com","logo":{"@type":"
          ImageObject","url":"https://images2.minutemediacdn.com/image
          /upload/c_fill,w_1440,ar_1:1,f_auto,q_auto,g_auto/shape/
          cover/sport/Favicon_90min-a4ca6e90ac5903d13cba99f629b6f1f1.
          png"}},"articleSection":"Transferências"} (...)
11    (...)
```

Código 2.1: Código HTML do *lobby* da página 160 do site 90min (2025).

Com a visualização do código, o objetivo é identificar onde estão os endereços digitais de cada uma das notícias apresentadas nesta parte do site. Após a verificação, identificamos que eles se encontram na linha 10 do Código 2.1 (neste código é apresentado apenas um endereço, mas os demais estão na sequência deste exemplo, na mesma *tag*), ou seja, estão descritos na *tag* *<script type="application/ld+json>"*, em um dicionário junto com outros metadados, cujo nome é "url".

2.2.2 Coletando os links

Com a identificação do local das URL's, por meio da *tag* detectada na seção anterior, o objetivo seguinte é o armazenamento e gerenciamento delas de forma organizada, permitindo seu acesso automatizado durante a coleta de dados. Com o auxílio das bibliotecas *request* e *bs4* (Richardson (2025)) realizaremos este passo.

```
import request as rq
from bs4 import BeautifulSoup
# Adicionando um usuário para evitar bloqueio do site
headers = {
"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit
    /537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"
}
# Tempo de atraso para acesso
time.sleep(random.uniform(0, 1))
# Acessando com no máximo 4 segundo de pesquisa
http = rq.get("https://www.90min.com/pt-BR/noticias-futebol?page=160",
    headers = headers, timeout = 4)
# Transformação em html
soup = BeautifulSoup(http.text, 'html.parser')
# Encontrando onde estão os links
urls = soup.find_all('script', type="application/ld+json")
```

Código 2.2: Captura do local das URL's do *lobby* da página 160.

Neste Código 2.2, *headers* será um usuário artificial para os inúmeros acessos que serão feitos, evitando o bloqueio de acesso ao site. Além disso, esse bloqueio também pode ocorrer por acessos sucessivos, portanto, há um tempo de atraso aleatório para evitar o bloqueio. O objeto *http* será a resposta HTTP da busca ao site solicitado, sendo necessário o usuário artificial (*headers*) e o tempo de leitura (*timeout*), em segundos, que o usuário levará para ler o conteúdo da página. Esse tempo foi definido após testes, pois tempos menores que 4 segundos não são suficientes para a leitura das páginas. Logo após, há a transformação desse objeto em HTML através da função *BeautifulSoup*, retornando ao formato de construção original da página. Por fim, com o código da página em mãos, selecionamos a *tag*, definida anteriormente no Código 2.1, onde estão localizadas as URL's.

Por estes passos coletamos a linha do código fonte da página que contém todos os URL's das notícias, neste exemplo, feito sobre o *lobby* da página 160 da plataforma. Logo, basta coletar tais endereços.

```
for url in urls:
    # Transformando cada local dos links em texto
    text = url.get_text()
    # Transformando em dicionários
    data = json.loads(text)
    # Acessando a chave em que está o dicionário
    json_data = data['itemListElement']
    # Para cada item do dicionário retiramos o link
    for item in json_data:
        links.append(item['url'])
```

Código 2.3: Captura das URL's do 160º *lobby*.

Portanto, do objeto *urls* temos um dicionário para cada notícia, como visto no Código 2.1. Logo, teremos a transformação de toda linha coletada para um dicionário no formato JSON, um dicionário universal, com uma das chaves sendo uma lista de JSON's, onde estão localizadas as URL's (Código 2.3). Com isso, acessando a chave *itemListElement* desse JSON em que estão os JSON's alvo(Código 2.4), obtemos uma lista de dicionários (*json_data*). Desta forma, acessando a chave *url* obtemos a URL de cada notícia pertencente a este *lobby*.


```
{
 '@type': 'NewsArticle',
 'position': 14,
 'url': 'https://www.90min.com/pt-br/corinthians-avisa-flamengo-
    prepara-compra-goleiro-hugo-souza',
 'headline': 'Corinthians avisa Flamengo e exerce compra do goleiro
     Hugo Souza',
 'image': 'https://images2.minutemediacdn.com/image/upload/(...)',
 'datePublished': '2024-10-10T18:01:07Z',
 'author': {'@type': 'Person',
     'name': 'FabrA\xadcio Carvalho',
     'url': 'https://www.90min.com/pt-BR/authors/fabricio-ca'},
 'publisher': {'@type': 'Organization',
     'name': '90min.com',
     'logo': {'@type': 'ImageObject',
      'url': 'https://images2.minutemediacdn.com/image/upload/(...)
```

```
            '}},
 'articleSection': 'Transferencias'
}
```

Código 2.4: Exemplo de uma notícia no *lobby* em formato JSON.

2.2.3 Acesso à notícia

O acesso à notícia acontece de maneira similar ao acesso ao *lobby*, porém utilizando o URL coletado na seção anterior. Após a transformação em HTML, basta encontrar a localização dos metadados a serem coletados por meio da *tag* utilizada, representada no Código 2.5.

```
    # Seleciona apenas o texto
    text = soup.find_all("p", class_ = "tagStyle_z4kqwb-o_0-
        style_1tcxgp3-o_0-style_1pinbx1-o_0-style_48hmcm")
    clean_text = "".join([t.get_text() for t in text])
    # Criando dicionário para a notícia
    conteudo = {}
    conteudo["Titulo"] = soup.find('title').get_text() # Titulo
    conteudo["Texto"] = clean_text # Texto
    conteudo["Link"] = link # Link
    conteudo["Data"] = soup.time["datetime"].split('T')[0] # Data
    author = json.loads(soup.find_all('script', type = 'application/ld+
        json')[0].get_text())['author'] # Autor
    if len(author) != 2:
        conteudo["Autor"] = author['name']
    else:
        conteudo["Autor"] = author[0]['name']
```

Código 2.5: Coleta dos metadados das notícias.

Portanto, a coleta dos metadados definidos anteriormente é realizada de acordo com a particularidade que são codificados e são adicionados a um novo dicionário em formato JSON. O objeto *soup* contém a codificação da página em HTML e, por meio das *tags*, localizamos os metadados da notícia.

Para a coleta do título é necessário encontrar a *tag title*. Já para o texto identificamos a *tag* descrita na definição do objeto *text*, onde se encontra o corpo da notícia e coletamos apenas o texto, pois as páginas digitais também são compostas de propagandas e outros

conteúdos textuais indesejados para este trabalho. O link é coletado de maneira automática, já que o detectamos anteriormente e a data é facilmente encontrada no código em *soup*, requerendo apenas uma transformação. Por fim, para o autor, após encontrar a *tag* no código, basta selecionar a chave "author" no JSON. Observamos que caso haja mais de um autor para a notícia, o primeiro nome será selecionado para compor os dados. Cada um destes dicionários será salvo separadamente para a conclusão deste processo com o próximo passo.

2.2.4 Banco de Dados final

Com o grande número de notícias, o número de arquivos JSON também seria alto e de difícil locomoção dentro do ambiente de trabalho. Com isso, a construção de um Banco de Dados torna-se ideal.

```
for file in full_files:
    # Abrindo cada arquivo json
    with open(file, 'r') as f:
        data = json.load(f)
        data['Titulo'] = data['Titulo'].encode('latin1').decode('utf-8')
        data['Texto'] = data['Texto'].encode('latin1').decode('utf-8')
        data['Autor'] = data['Autor'].encode('latin1').decode('utf-8')
    # Adicionando na lista criada
    if isinstance(data, list):
        full_data.extend(data)
    else:
        full_data.append(data)
    (...)
with open(csv_file / "Data_news.csv", 'w', newline='', encoding='utf-8')
    as csv_out:
    # Selecionando o nome das colunas
    headers = full_data[0].keys()
    # Criando o local para adicionar os textos
    writer = csv.DictWriter(csv_out, fieldnames=headers)
    # Escrevendo o nome das colunas
    writer.writeheader()
    # Escrevendo os conteudo
    writer.writerows(full_data)
```

Código 2.6: Criação do Dataset de notícias.

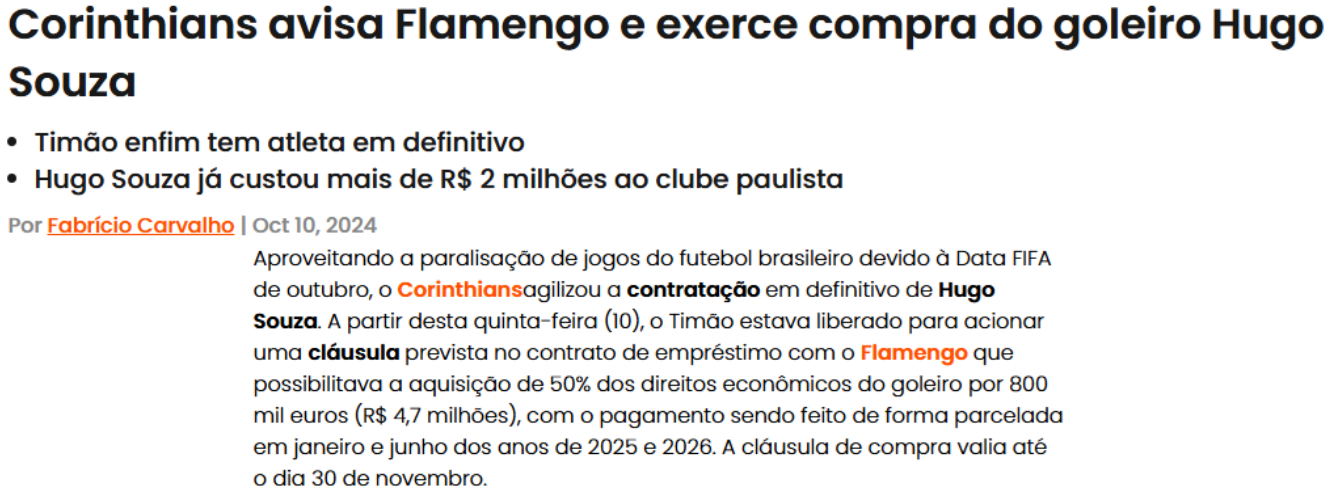

Figura 2.3: Notícia de exemplo na página digital.

Logo, cada uma das colunas deste banco será um metadado coletado da notícia; observamos que os metadados Título, Texto e Autor, que são textos, salvos no JSON estão escritos de acordo com a codificação em HTML, para isso usamos a combinação de funções *encode('latin1').decode('utf-8')* que transformará o texto para a língua portuguesa. Em seguida, adicionamos os JSON's em uma lista para que, ao criar um arquivo no formato CSV possamos nomear as colunas com as chaves do dicionário e escrever tais metadados em sua respectiva coluna por meio das funções das linhas 20 e 22 do Código 2.6. Por fim, para resumo do processo, apresentamos uma notícia, que foi utilizada de exemplo nesta seção.

A Figura 2.3 retrata como temos a notícia na página digital, mostrando todos os metadados a serem coletados. Em seguida, no Código 2.7 temos o resultado da coleta dos metadados.

```
1 {
2     "Título": "Corinthians avisa Flamengo e encaminha compra do
          goleiro Hugo Souza",
3     "Texto": "Aproveitando a paralisação de jogos do futebol
          brasileiro devido à Data FIFA de outubro, o Corinthians
          agilizou a contratação em definitivo de Hugo Souza. A partir
           desta quinta-feira (10), o Timão (...)",
4     "Link": "www.90min.com/pt-BR/corinthians-avisa-flamengo-prepara
          -compra-goleiro-hugo-souza",
```

```
    "Data": "2024-10-10",
    "Autor": "Fabrício Carvalho"
}
```

Código 2.7: Notícia de exemplo em JSON.

Portanto, o processo de coleta resultou em um Banco de Dados com cerca de 8.000 observações, distribuídas entre outubro de 2023 e setembro de 2025, sendo uma observação equivalente a uma notícia e suas variáveis análogas aos metadados. Esse período foi escolhido por cobrir o fim da temporada de 2023, todo o ano de 2024 e a temporada mais atual de 2025, permitindo acompanhar os desdobramentos dos acontecimentos do ano anterior. As publicações abordam temas diversos do mundo futebolístico, como resultados de jogos, negociações e transferências de jogadores (Mercado da Bola), entrevistas, declarações públicas e também polêmicas envolvendo profissionais do futebol. Essa variedade de assuntos fornece uma base sólida para a realização das análises neste trabalho. Na Tabela 2.1 temos alguns exemplos de como o Banco de Dados está estruturado para esta monografia.

Tabela 2.1: Exemplo de observações no banco de dados final.

Título	Texto	Link	Data	Autor
Corinthians avisa Flamengo e encaminha compra do goleiro Hugo Souza.	Aproveitando a paralisação de jogos do futebol brasileiro devido à Data FIFA de outubro, o Corinthians (...)	https://www.90min.com/pt-BR/corinthians-avisa-flamengo-prepara-compra-goleiro-hugo-souza	2024-10-10	Fabrício Carvalho
Corinthians abre negociações para contratar Balotelli.	Em busca de um nome de impacto para marcar o primeiro ano da gestão no Corinthians, o presidente Augusto Melo (...)	https://www.90min.com/pt-BR/posts/corinthians-abre-negociacoes-contratar-mario-balotelli	2024-07-09	Fabrício Carvalho

Capítulo 3

Processamento

Neste capítulo, serão apresentados os principais conceitos teóricos e práticos das etapas deste trabalho. A análise de textos em larga escala, como os extraídos de portais de notícias, exige uma série de transformações e interpretações computacionais para que o conteúdo textual possa ser compreendido e utilizado por algoritmos para a geração de resultados adequados. A combinação destas técnicas permite que modelos computacionais lidem com grandes quantidades de textos de maneira eficiente e significativa, tornando possível uma análise aprofundada.

3.1 Pré-processamento

Há certas palavras contidas nos textos que servirão apenas para o cumprimento de regras gramaticais, não gerando, portanto, um significado relevante para o entendimento do leitor dos mesmos. Esses termos, em PLN, não agregarão informação que possa ser digna de análise, justamente por não terem consigo um valor que alterará o significado da frase ou texto que estamos analisando. Logo, certas classes de palavras são caracterizadas desta forma: artigos, conjunções, conectores e outras classes gramaticais que apenas auxiliam na construção dos textos. Manning *et al.* (2008), portanto, nomeiam tais palavras como *stopwords* e serão excluídas dos próximos passos do processo.

Ademais, ao selecionar diversos textos é provável que tenhamos uma palavra que tem uma frequência maior em todos eles, algum verbo ou adjetivo a depender da característica do gênero textual. Nas notícias de futebol, por exemplo, esperamos que algumas palavras possuam uma frequência muito grande entre elas, tais como "bola", "clube" e "competição" que não pertencem à mesma classe gramatical das citadas anteriormente. Portanto,

também podemos caracterizar tais palavras como *stopwords*, pois elas não agregarão grandes resultados por seu uso frequente no contexto futebolístico. Neste trabalho, usaremos o dicionário de *stopwords* pré-definidas pela biblioteca em python NLTK (Aarsen *et al.* (2024)) e também definiremos outras que são possíveis de serem vistas apenas ao observar as suas frequências e as avaliando como não proveitosas para as análises dentro do contexto futebolístico; tal avaliação será feita posteriormente nas análises descritivas.

Definido, o que talvez seja, o principal conceito do pré-processamento dos textos, vamos introduzir outras boas práticas que visam a maior eficiência de PLN.

- **Acentuação:** em alguns idiomas, como o português, certas palavras possuem letras que são acentuadas. Til, crase, circunflexo e agudo são os acentos da língua portuguesa, além do cedilha, que podem gerar desentendimento nas análises, pois serão analisadas separadamente palavras iguais, mas distintas somente por erros de escrita. Por exemplo: "campeão" e "campeao";

- **Pontuação:** diferente da acentuação, as pontuações não mudarão significativamente as análises. Porém, se tratando de algoritmos de PLN simplistas (que retornam frequências e similaridades entre palavras), eles não serão capazes de diferenciar uma pergunta de uma afirmação e também mudar a análise por causa de uma vírgula, por exemplo. Portanto, retirar as pontuações pode beneficiar a rapidez do código, por não armazenar informações improdutivas;

- **Letras capitais:** semelhante as acentuações, padronizar todo o texto sem letras maiúsculas pode gerar um ganho no momento das análises. Nomes e palavras no começo das frases normalmente são escritas com a primeira letra sendo maiúscula, o que impede os agrupamentos dos mesmos nomes escritos de maneira indevida, como "Neymar" e "neymar", e de palavras no começo e no meio da frase.

- **Números:** dentro deste projeto, analisar números dentro dos textos não será o foco. Por isso, qualquer número que estiver contido no texto, como datas e valores de negociações, será removido.

```
1 {
2     Aproveitando a paralisação de jogos do futebol brasileiro
3     devido à Data FIFA de outubro, o Corinthiansagilizou a
4     contratação em definitivo de Hugo Souza. A partir desta
```

```
5      quinta-feira (10), o Timão (...)
6  }
```

Código 3.1: Texto antes do pré-processamento.

Logo, com o texto no Código 3.1 criamos o vetor de palavras do Código 3.2, sem palavras repetidas, se houverem.

```
1  ['aproveitando', 'paralisacao', 'jogos', 'futebol', 'brasileiro', '
       devido', 'data', 'fifa', 'outubro', 'corinthians', 'agilizou', '
       contratacao', 'definitivo', 'hugo', 'souza', 'partir', 'desta', '
       quinta', 'feira', 'timao', (...)]
```

Código 3.2: Texto após do pré-processamento.

Estas técnicas citadas foram adotadas para a análise dos textos desta monografia para que os resultados pudessem ser maximizados, evitando redundâncias e dissipação dos efeitos das mesmas palavras, porém escritas de maneiras distintas.

3.2 Vetorização

Com o primeiro passo do processo dado, vamos iniciar a etapa responsável por permitir que as análises dos textos sejam feitas. A vetorização das palavras é uma forma de padronizar cada termo dentro de um conjunto de textos, que possibilitará a visualização das palavras como números e, não apenas isso, mas com a vantagem de serem interpretáveis dentro do contexto analisado.

Esta transformação considera a posição da palavra de referência dentro do texto e as demais palavras que a acompanha. Eliminando as *stopwords*, termos com algum significado tendem a ficar próximos e, com isso, ao transformá-los em números também tendem a ter uma relação próxima nas análises. Na programação, um objeto no formato de vetor é muito usado por sua fácil leitura e manuseio. Em PLN, temos algumas técnicas de vetorização de palavras, as mais usadas são *Bag-of-Words* (Joachims (1998)) e *Word Embeddings* (Bengio *et al.* (2003)). Neste trabalho faremos uso da segunda técnica citada, *Word Embeddings*.

Word Embeddings faz uso de um dos algoritmos que revolucionou a área de Aprendizado de Máquina (AM) no século XX, as Redes Neurais (Haykin (1994)). As Redes Neurais Artificiais se consolidaram como uma das principais ferramentas da área de AM,

sendo capazes de modelar relações complexas entre os dados e de generalizar padrões mesmo em contextos de alta dimensionalidade. Logo, podemos entender que uma rede é formada por camadas compostas por neurônios artificiais que recebem entradas numéricas, aplicam uma transformação matemática e retornam um resultado, de acordo com o contexto em que é aplicada.

Em seu artigo, Bengio *et al.* (2003) citam como a vetorização pode ser difícil devido ao alto número de palavras que um vocabulário pode ter ao analisarmos um número considerável de textos. Portanto, este método visa, dentro de um espaço vetorial limitado, estimar as generalizações semânticas entre as palavras que compõem os textos, evitando assim a vetorização de conjuntos desnecessários. A limitação do espaço vetorial é feita ao determinar o tamanho do vetor que representará cada palavra, é importante que esse tamanho seja menor que o número de palavras presentes no texto (vocabulário) para que haja uma eficiência maior desta técnica.

Mikolov *et al.* (2013) foram responsáveis por aprimorar as técnicas de *Word Embeddings* com o desenvolvimento do modelo *Word2Vec*, possuindo uma boa eficiência na captura das relações semânticas e sintáticas das palavras. Ademais, o método pode ser definido por duas vertentes, *Skip-gram* e *Continuous Bag of Words* (CBOW), ambas apresentadas neste artigo. *Skip-gram* utiliza a palavra de referência para prever as palavras que se encaixam com ela, portanto, prevê o contexto que melhor se enquadra à palavra; o método CBOW usa a técnica contrária, por meio das palavras de uma frase fornecida, prevê a palavra que melhor se encaixa na frase.

Ambas metodologias fazem uso do método primário de vetorização, o *One-Hot*. Martin e Jurafsky (2025) introduzem ideias sobre a representação das palavras que serviram como base para o *One-hot encoding*. Com este objetivo, a vetorização cria um vetor de igual tamanho ao número de palavras da frase e cada palavra é atribuída a um vetor de zeros e um único 1 na posição em que a mesma se encontra na frase. Em suma, a técnica de *Skip-gram* será utilizada neste trabalho para a vetorização das palavras, por sua maior eficiência em um baixo número de dados.

A Figura 3.1 representa o processo de vetorização que será feito dentro da função *Word2Vec*. Para o método temos uma rede neural rasa, ou seja, que possui apenas uma camada oculta. Este fato auxilia no objetivo do processo, que é aprender representações vetoriais que preservem relações semânticas entre as palavras. Com isso, dado uma notícia, teremos o objeto *Texto*, posto na camada de Entrada para que a transformação *One-Hot*

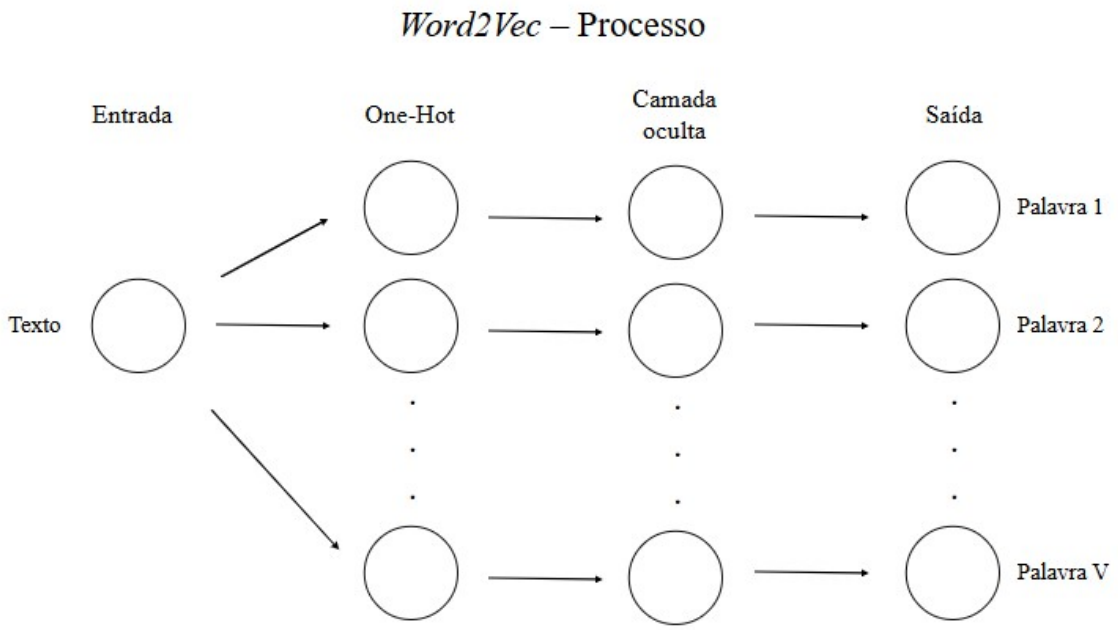

Figura 3.1: Processo da função *Word2Vec* para vetorização.

seja realizada; na sequência, na Camada oculta será feita a transformação das palavras com a técnica *Skip-gram* e com função de ativação *Softmax*. Por fim, em Saída estarão cada palavra do vocabulário do texto como vetores, sendo i o i-ésimo vetor, com $i = 1, 2, ..., V$, $V \in \mathbb{N}$.

Em Python este modelo está contido na biblioteca *gensim* (Rehurek (2024)) e a função leva o mesmo nome do modelo. Certos argumentos dessa função são relevantes para as análises por serem os hiper-parâmetros do modelo em questão; eles serão descritos a seguir.

- *sentences:* entrada para os textos a serem considerados na estimação. Neste caso, será introduzida uma lista - com cada índice sendo um texto de listas - com um texto sendo apresentado por suas palavras em cada índice, após o pré-processamento;

- *vector_size:* argumento que define o tamanho do vetor que representará cada palavra; é responsável por limitar o espaço vetorial;

- *window:* fixando a palavra a ser estimada, é o número de palavras antes e depois da referência; um intervalo que aumenta a relevância das palavras que estão contidas nele;

- *min_count:* frequência mínima que cada palavra deve ter em *sentences*; caso contrário é retirada da estimação.

O número de notícias disponíveis no Banco de Dados é considerada baixa por estarmos trabalhando dentro do contexto de PLN, sabendo da gama de palavras que são pertinentes dentro do tema proposto. Logo, os argumentos *vector_size* e *window* são sensíveis a grandes variações. Para o primeiro argumento, um vetor de grande dimensão pode ser complexo para o problema tratado e o contrário pode generalizar as relações entre termos. Já para o segundo argumento, intervalos grandes podem conter palavras que estão distantes à de referência e, portanto, palavras sem concordância (dentro do contexto) passariam a ter maiores relevâncias na estimação, algo que não queremos, e com intervalos pequenos, somente termos muito próximos serão considerados, podendo descartar similaridades importantes.

Uma das características deste método, além de trabalhar com as palavras de maneira vetorial, é a possibilidade de escrever equações matemáticas com os vetores. Isso permite encontrar relações dentro das notícias que dizem respeito às particularidades do contexto em que as palavras estão inseridas. Por exemplo, aplicando este método, a equação entre as palavras,

$$Barcelona - Europa + Brasil,$$

obteve como resultado um vetor que tinha como palavra similar o termo "Palmeiras". Logo, mesmo sem compreender o significado das palavras que estão na operação matemática, o algoritmo consegue resultar em interpretações próximas à realidade, apenas observando as posições em que as palavras se encontram nos textos. Para o exemplo citado, "Barcelona" se refere a um time da Europa e, ao subtrair a palavra "Europa" e somar a "Brasil" , o resultado se assemelha ao clube do Brasil, "Palmeiras" . Na Figura 3.2 estão representados estes vetores, vetorizados em 20 dimensões, pelo método citado, e reduzidos a 2 dimensões pelo método de Componentes Principais, para que a exposição gráfica fosse possível.

Ademais, da mesma maneira que o algoritmo é capaz de interpretar palavras próximas dentro dos textos, indicar termos que são distintos dentro do contexto também é um dos cenários possíveis de ser aplicado. Esta propriedade pode ser usada para indicar possíveis anomalias dentro do texto, com resultados que não expressam corretamente a realidade. No contexto futebolístico, sabemos que clubes de futebol se diferenciam de clubes brasileiros em diversos aspectos. Ao utilizar desta técnica com as palavras "Corinthians", "Flamengo" e "Liverpool", temos que o retorno de anomalia dentre estes

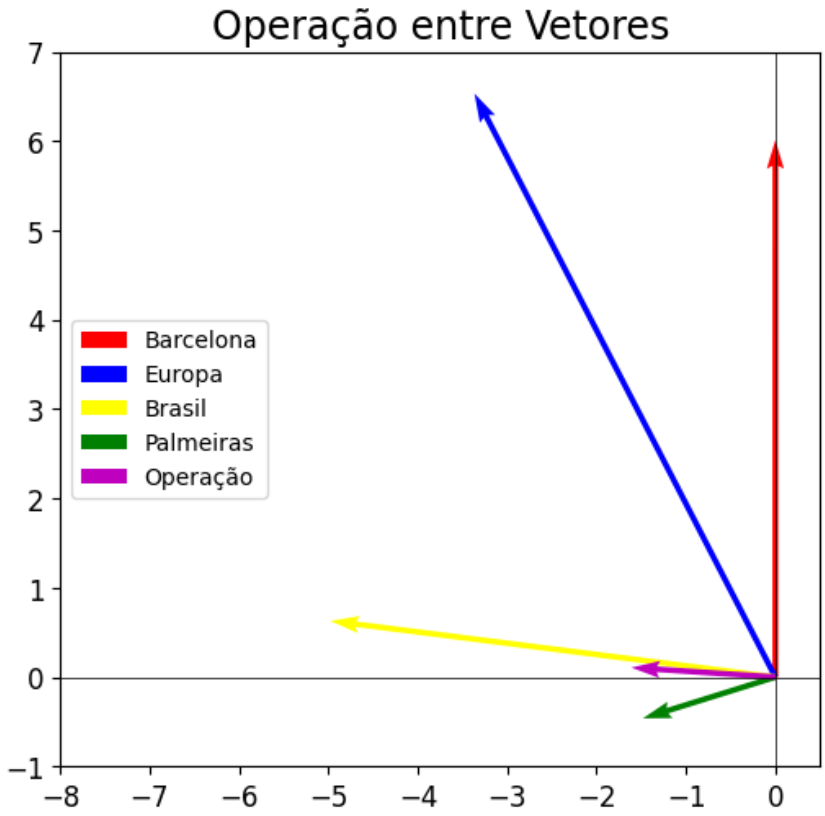

Figura 3.2: Exemplo da operação matemática entre palavras.

se dá pelo clube da Inglaterra, Liverpool, afinal os outros dois são clubes sediados no Brasil.

3.3 Análise de Similaridade

Com o método descrito anteriormente, as palavras agora serão analisadas como vetores, porém mantendo as relações semânticas e sintáticas das palavras. Salton *et al.* (1975) introduziram o cálculo da distância entre vetores gerados para recuperação de informação. Mais tarde, seu uso foi expandido para diversas área, incluindo a de PLN. Nesse contexto, portanto, usaremos este cálculo para averiguar o quão similares os vetores são e, consequentemente, qual o nível de relacionamento das palavras dentro do texto.

Portanto, sejam $\mathbf{u} = (u_1, ..., u_n)$ e $\mathbf{v} = (v_1, ..., v_n)$ vetores pertencentes a $\mathbb{R}^n$. Definimos o cosseno do ângulo entre os vetores como,

$$cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|},$$

onde a similaridade entre os vetores é o próprio valor do cosseno, $\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i$,

$\|\mathbf{u}\| = \sqrt{\sum_{i=1}^{n} u_i^2}$, $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}$ e θ o ângulo formado entre os dois vetores no espaço vetorial assumido.

O cosseno retorna valores no intervalo $[-1, 1]$, porém, devido à metodologia usada neste trabalho (representação dos vetores em altas dimensões), palavras similares tendem a serem paralelas e não similares a serem ortogonais. Portanto, os valores de similaridade variam de $[0, 1]$, em que 0 são palavras que tendem a não ter relação alguma e 1 sendo palavras que tendem a estarem relacionadas dentro do texto. Essa característica permite analisar o comportamento das relações entre as palavras dos textos observados, permitindo a interpretação das mesmas dentro do contexto futebolístico.

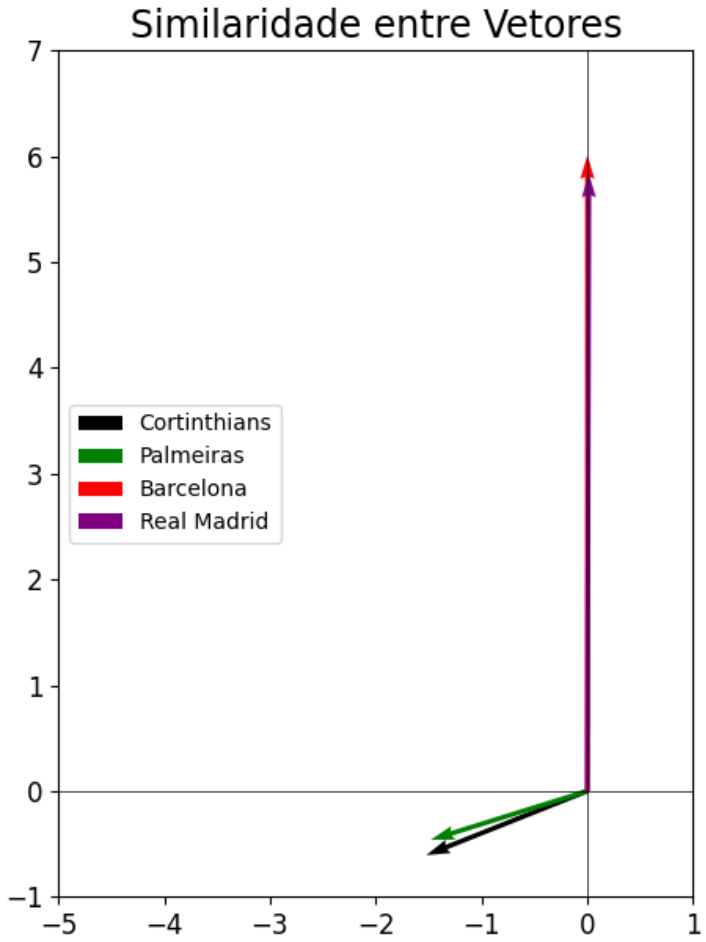

Figura 3.3: Exemplo de similaridades entre as palavras.

Como exemplo, a Figura 3.3 apresenta quatro palavras dos textos que, vetorizadas em 20 dimensões, pelo método *Word2Vec*, e reduzidas a 2 dimensões, com a técnica de Análise de Componente Principais, que será descrita posteriormente, são representadas pelos vetores destacados. Vemos que, para as palavras "Corinthians" e "Palmeiras", os seus respectivos vetores possuem um ângulo menor que entre as palavras "Barcelona" e

"Palmeiras" e, portanto, as similaridades destes vetores são distintas. Para "Corinthians" e "Palmeiras" a similaridade é de 0.997, entre "Barcelona" e "Real Madrid" , de 1 e, para os vetores mais próximos destas duplas, "Palmeiras" e "Barcelona" , de -0.372. Portanto, indicando que as primeiras palavras citadas são mais similares que o par "Barcelona" e "Palmeiras" que, no contexto futebolístico, é razoável tal indicativo, uma vez que o clube Palmeiras é sediado no Brasil e não teria relação, a priori, com o clube espanhol Barcelona e a palavra "Corinthians" sendo o principal rival deste clube no meio do futebol. Além disso, Barcelona e Real Madrid também formam a principal rivalidade entre clubes da Espanha. É importante ressaltar que tais similaridades foram calculadas considerando os vetores no espaço de duas dimensões e, por terem "menos espaço", quando comparados aos mesmos vetores no espaço de dimensão 20, acabam resultando em vetores de palavras muito similares (com similaridades muito próximas a 1) ou de similaridades negativas.

3.4 Identificação de Entidades

A identificação de entidades dentro de um texto pode facilitar as análises de PLN, em uma visão macro do contexto. Definimos entidades como sendo nomes próprios, locais, organizações e até mesmo, a depender do objetivo, datas e obras contidas no texto. Grishman e Sundheim (1996) introduziram o conceito de NER (*Named Entity Recognition*) como uma técnica inicial para as análise de PLN que, por meio dessa tarefa central, define as categorias das palavras e avalia os resultados com métricas estatísticas.

O método auxilia o pesquisador a encontrar palavras de interesse dentro do próprio vocabulário. Como temos por objetivo analisar os fatos que ocorrem no futebol, coletar os nomes de jogadores, técnicos, clubes, estádios e outras classes de palavras do meio futebolístico é fundamental para a análise. Com o número de notícias obtido, a coleta manual destas classes é difícil e o uso dessa técnica aumenta a eficiência do processo.

Em programação, existem alguns modelos já construídos para aplicação do método. Neste trabalho faremos uso da biblioteca *spaCy* (Honnibal (2024)) que possui um modelo pré-treinado para a coleta das entidades e aplicação simplificada dessas regras.

```
# Definindo as regras para pegar as entidades de acordo com os textos
reader = sp.load('pt_core_news_sm')
entites = text_n.apply(lambda x: reader(x).ents)

# Visualizar as entidades e seus rótulos em News
```

```
all_entites = []
all_labels = []
for e in entites:
    for i in range(len(e)):
        if np.isin(e[i].text, all_entites) == False:
            all_entites.append(e[i].text)
            all_labels.append(e[i].label_)
```

Código 3.3: Aplicação do NER nos textos.

No Código 3.3, apresentamos a aplicação da técnica usando o modelo pré-treinado para a língua portuguesa. Este modelo faz uso de Redes Neurais Convolucionais (LeCun *et al.* (1998)) para identificação das entidades e suas classificações, após passar por um processo de vetorização das palavras. E, após isso, há uma verificação para evitar a adição de entidades repetidas.

Este código retornará tais entidades e suas categorias encontradas dentro dos textos fornecidos, como:

- PER: nomes próprios de jogadores, técnicos e outros profissionais do meio futebolístico;
- LOC: localidades que vão desde cidades até estádios;
- ORG: organizações, referentes aos clubes, neste contexto;
- MISC: demais entidades sem classificação específica.

Logo, como não faremos uso da classe MISC, a retiramos do conjunto de entidades. Portanto, temos um conjunto de palavras que foi coletado por este algoritmo que servirá de base para as futuras análises acerca de jogadores, clubes e personagens do meio futebolístico desta monografia.

3.5 Redução de Dimensionalidade

A Redução de Dimensionalidade é uma técnica estatística utilizada para, de alguma maneira, resumir as informações contidas em um banco de dados. Neste contexto, o objetivo da aplicação é identificar palavras que possuem um mesmo comportamento nos textos e apresentar, graficamente, os agrupamentos resultantes. Desta forma, a assimilação e identificação de palavras relacionadas será de fácil acesso.

3.5.1 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) é um método desenvolvido inicialmente por Pearson (1901) e generalizado por Hotelling (1933). Trata-se de uma técnica matemática que busca transformar um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de variáveis não correlacionadas, os componentes principais. Cada componente é formado por uma combinação linear das variáveis originais e é construído de modo a capturar a maior variância possível dos dados.

Portanto, dado um conjunto de dados $X \in \mathbb{R}^{n \times p}$, com n observações e p variáveis, a análise encontra uma matriz de projeção $W \in \mathbb{R}^{p \times k}$,

$$W = \begin{bmatrix} \mathbf{w_1} & \mathbf{w_2} & \cdots & \mathbf{w_k} \end{bmatrix},$$

tal que,

$$\Sigma \mathbf{w_i} = \lambda_i \mathbf{w_i},$$

onde, Σ é a matriz de covariância de X e λ_i seus autovalores, $\mathbf{w_i}$ é o i-ésimo autovetor da matriz Σ, sendo W, portanto, a matriz de autovetores da matriz Σ. Com isso, encontramos a igualdade,

$$Z = XW,$$

em que Z, $Z \in \mathbb{R}^{n \times k}$, representa os dados projetados em um espaço de dimensão reduzida k, com $k < p$. Desta maneira, a variância dos dados se concentra no primeiro componente principal, no segundo se concentra a maior parte da variância restante, e assim sucessivamente, desde que os demais componentes sejam ortogonais aos primeiros. De forma prática, as palavras serão representadas nas linhas de X como vetores transpostos de dimensão $1 \times p$, com $p \in \mathbb{N}$, sendo o número de dimensões definido na vetorização. Com isso, em W estarão, nas colunas, os autovetores de dimensão $p \times 1$ e, para o objetivo do estudo, que é realizar uma representação gráfica dos vetores das palavras, $k = 2$.

Com isso, a ACP permite reduzir a dimensionalidade de um conjunto de dados preservando a maior parte da informação. Isso facilita a análise, a interpretação e a representação gráfica de variáveis em espaços de dimensão reduzida. Portanto, a ACP permite a projeção dos vetores de palavras em um espaço bidimensional, possibilitando a visualização gráfica das relações entre os termos.

3.5.2 t-SNE

O t-SNE (*t-distributed Stochastic Neighbor Embedding*) é um método de redução de dimensionalidade proposto por van der Maaten e Hinton (2008). Ele foi desenvolvido com o objetivo de projetar dados de alta dimensão em espaços de duas ou três dimensões, preservando as relações de vizinhança entre os pontos.

A ideia central do t-SNE é converter distâncias euclidianas entre pontos em distribuições de probabilidades que representam similaridades. Em alta dimensão, a similaridade entre dois pontos x_i e x_j é definida como a probabilidade condicional de que x_j seria escolhido como vizinho de x_i, considerando uma distribuição gaussiana centrada em x_i,

$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)},$$

simetrizando essas probabilidades,

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

Já no espaço de baixa dimensão, com pontos y_i e y_j, as similaridades são modeladas por uma distribuição proporcional a t-Student, com um grau de liberdade, definida por,

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}.$$

Portanto, o algoritmo busca encontrar a projeção em baixa dimensão que minimize a divergência de Kullback e Leibler (1951) entre as distribuições de probabilidade das similaridades descritas, $P = \{p_{ij}\}$ e $Q = \{q_{ij}\}$,

$$C = KL(P \,\|\, Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Logo, esse processo faz que pontos semelhantes no espaço original permaneçam próximos no espaço reduzido, enquanto pontos distantes tendem a ser separados. Isso permite que padrões complexos, como agrupamentos semânticos de palavras, sejam visualizados de maneira clara, mesmo quando a estrutura dos dados originais é não linear. Por exemplo, palavras relacionadas a futebol tendem a se agrupar em regiões próximas do espaço projetado, separadas de termos de outros contextos.

Capítulo 4

Resultados

Neste capítulo, são apresentados e discutidos os principais resultados obtidos ao longo do desenvolvimento deste trabalho. Após a preparação dos dados e aplicação dos conceitos apresentados, os produtos gerados por essas análises serão expostos com o objetivo de demonstrar como os fatos ocorridos no período de estudo refletem nas análises. Além da exposição dos dados, este capítulo busca interpretar as relações encontradas, evidenciando tendências, comportamentos e possíveis implicações. Portanto, os resultados retratam a combinação das técnicas estatísticas e de PLN citadas nos capítulos anteriores.

4.1 Análise Descritiva

A análise descritiva dos dados terá por objetivo compreender o número de notícias, de maneira geral ou individualizada para determinado termo. Logo, o entendimento de como as variações das frequências para diferentes palavras ocorrem servirá para as interpretações realizadas posteriormente.

Inicialmente, o número de notícias coletadas é insuficiente para uma análise com resultados consistentes em intervalos de tempo curtos, como diário, semanal e mensal. Logo, o agrupamento delas por trimestre foi necessário para que haja uma número de observações ideal para tais análises. Portanto, as investigações serão para conclusões sobre os comportamentos das palavras em um intervalo de tempo trimestral, sendo segmentados de maneira que Outubro a Dezembro de 2023 seja o primeiro trimestre, Janeiro a Março de 2024 seja o segundo e assim sucessivamente, até o último trimestre, Julho a Setembro de 2025. Com tal agrupamento, a Figura 4.1 descreve o resultado da frequência de notícias para cada trimestre e vemos um comportamento uniforme entre os períodos.

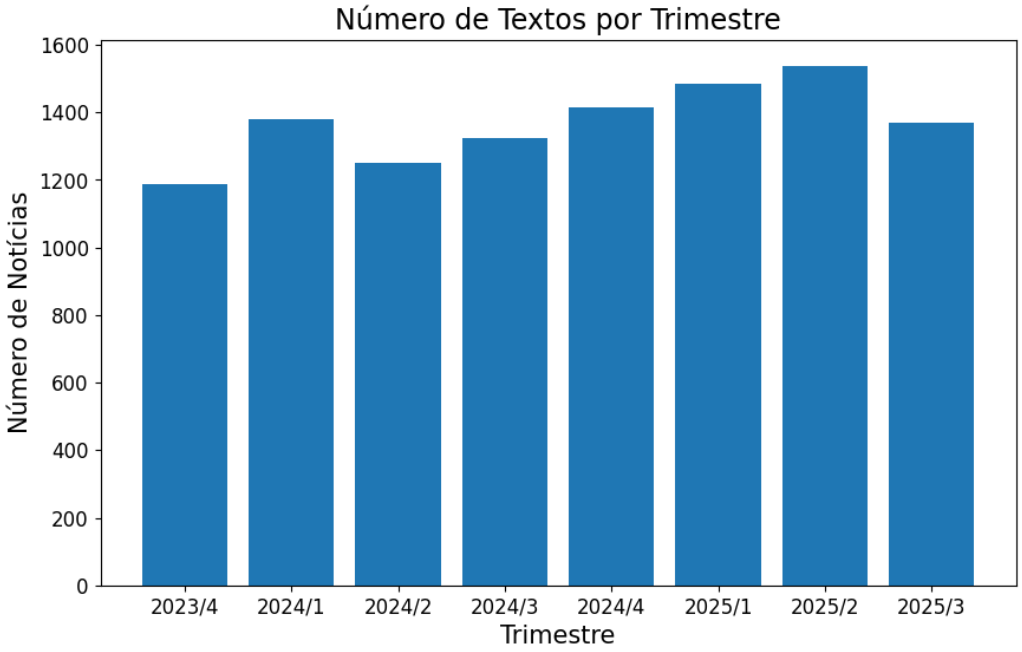

Figura 4.1: Número de notícias por trimestre.

Em seguida, a escolha de entidades, encontradas pelo processo descrito no Capítulo 3, que possuíam uma frequência total de ao menos 80 citações, para entidades PER e LOC, e 60 citações, para entidades ORG, em diferentes notícias foi feito para que, dadas as entidades filtradas, os resultados fossem consistentes. Desta forma a palavra de referência terá um número suficiente de palavras vizinhas, aumentando sua variabilidade, não concentrando as suas relações em poucos termos devido ao fato da referência ter poucos vizinhos, mas sim para que essa concentração exista somente quando um termo realmente tiver concordância com a referência.

Além da decisão acima, observamos na Figura 4.2 que, para as entidades PER, como citado no Capítulo 3, há uma análise inconclusiva sobre quais personagens do meio futebolístico estamos lidando. Isso ocorre pois, além do fato de diversos jogadores terem o primeiro nome igual, a citação de seus nomes nos textos, que falam separadamente sobre os mesmos, também pode ser feita apenas com o seu primeiro nome, algo que o algoritmo não compreende, afinal a identificação é feita em apenas um corpo formado por todas as notícias. Logo, uma análise geral das entidades torna-se inviável se os termos mais comuns fossem considerados, culminando para uma análise individual de um número limitado de entidades escolhidas pelo autor.

A palavra selecionada para análise, dentro da categoria PER, será "Neymar". Aqui, temos por objetivo analisar como o nome do jogador foi vinculado às notícias durante o

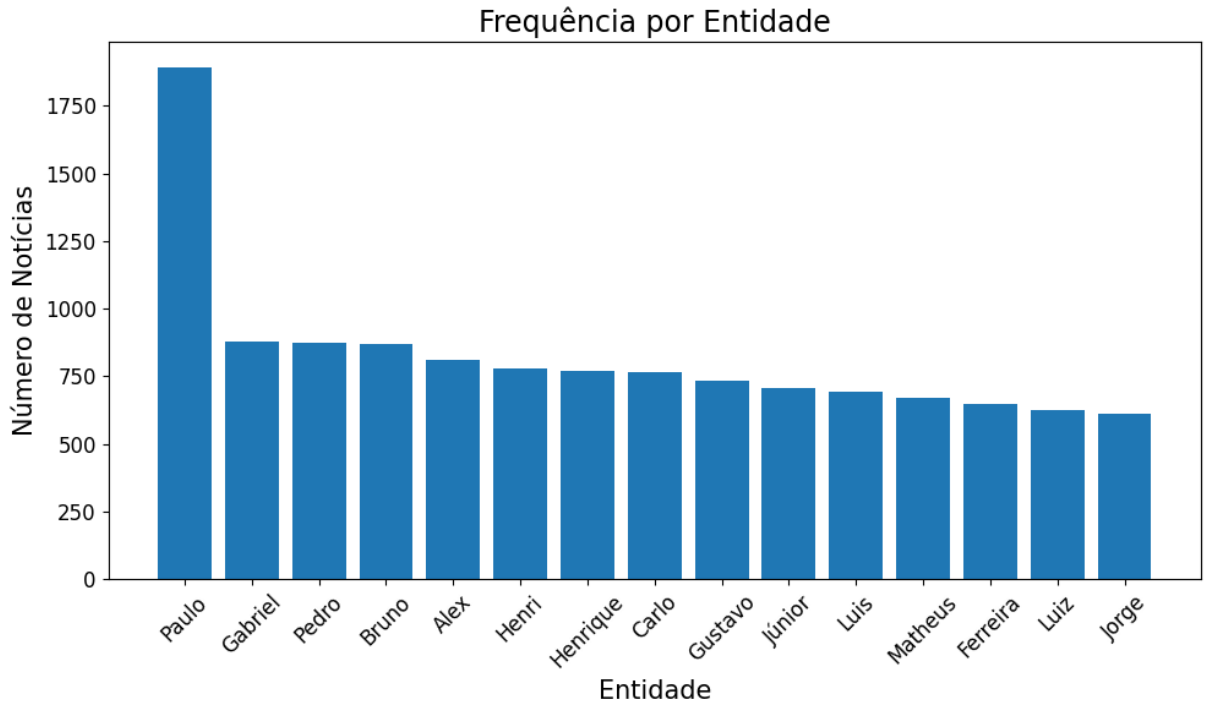

Figura 4.2: Entidades mais citadas.

período de estudo. Em particular, este profissional por vezes também é nomeado como "Ney", logo, houve a necessidade de uma transformação nos textos para que as duas entidades identificadas, "Ney" e "Neymar", fossem consideradas com o mesmo significado.

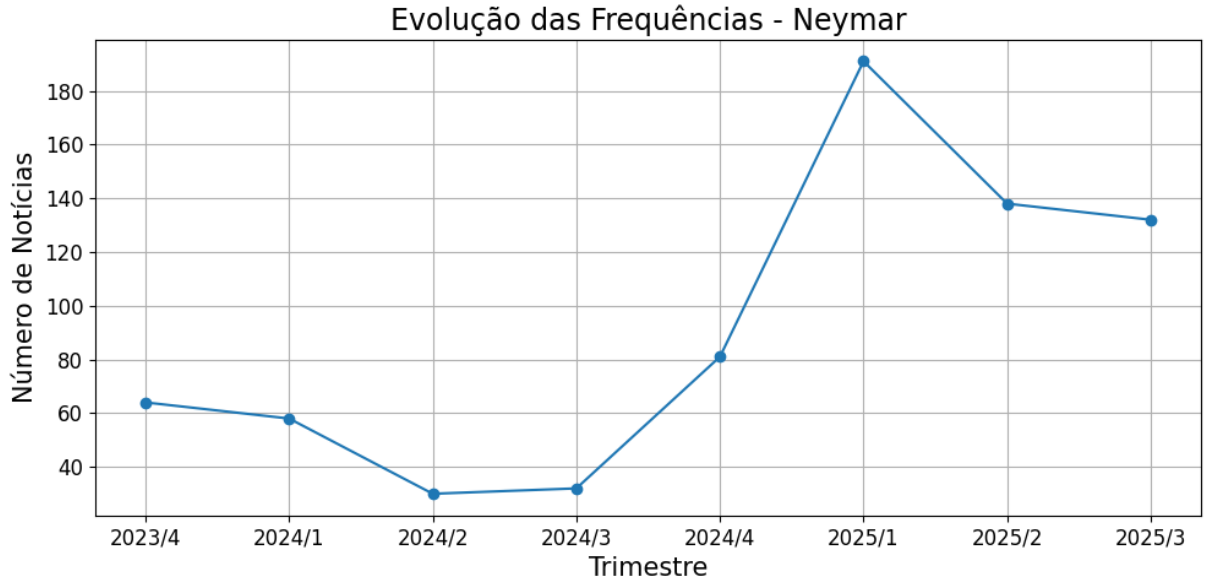

Figura 4.3: Número de notícias com a palavra "Neymar".

Na Figura 4.3 observamos que o número de notícias vinculadas ao jogador se manteve baixa durante o fim do ano de 2023 e durante todo ano de 2024. O fato do atleta ter convivido com lesões durante esse período indica a iminente queda midiática do mesmo

pela falta de aparições em jogos oficias pelo clube de futebol em que ele atuava na época. Além disso, a atuação do jogador em um clube pertencente a um campeonato alternativo, fora dos grandes centros, também contribui para tal queda. Mais adiante, vemos que em 2025 houve um grande aumento no número de citações, pois o atleta, nesse período, se transferiu para um clube brasileiro gerando um significativo aporte midiático e, consequentemente, evolução das notícias citando-o.

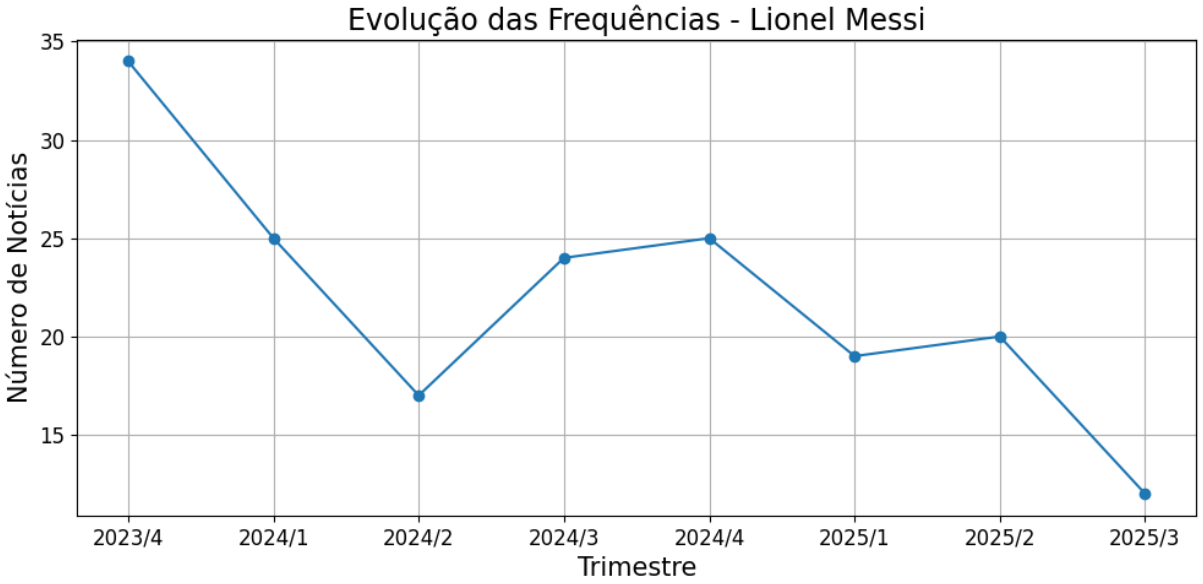

Figura 4.4: Número de notícias com a palavra "Lionel Messi".

Diante de uma outra entidade PER encontrada temos o jogador Lionel Messi. Em seu histórico, este profissional sempre esteve nos principais tabloides da mídia, por seus títulos, feitos e premiações. Porém, como visto na Figura 4.4, o número de notícias em que seu nome é citado decaiu durante o período analisado, o que nos fornece indicativos de que sua recente transferência para um Campeonato de Futebol alternativo, dentro do meio futebolístico, prejudicou suas aparições nos conteúdos midiáticos e diminuiu sua força publicitária nos noticiários.

Na Figura 4.5, identificamos as notícias que citaram a palavra "Liverpool", uma entidade LOC. Esta palavra representa um clube inglês de grande expressão em seu país. No gráfico, vemos a formação de dois picos, localizados nos períodos do começo de 2024 e fim de 2024 e começo de 2025. Nesses intervalos acontecem eventos chamados de "Janela de Transferência", dias em que os clubes podem trocar (vender ou comprar) jogadores do seu elenco para os campeonatos. Este clube, por sua vez, protagonizou rumores de transferências e contratações de jogadores que atuavam em grandes ligas, em grandes clubes e

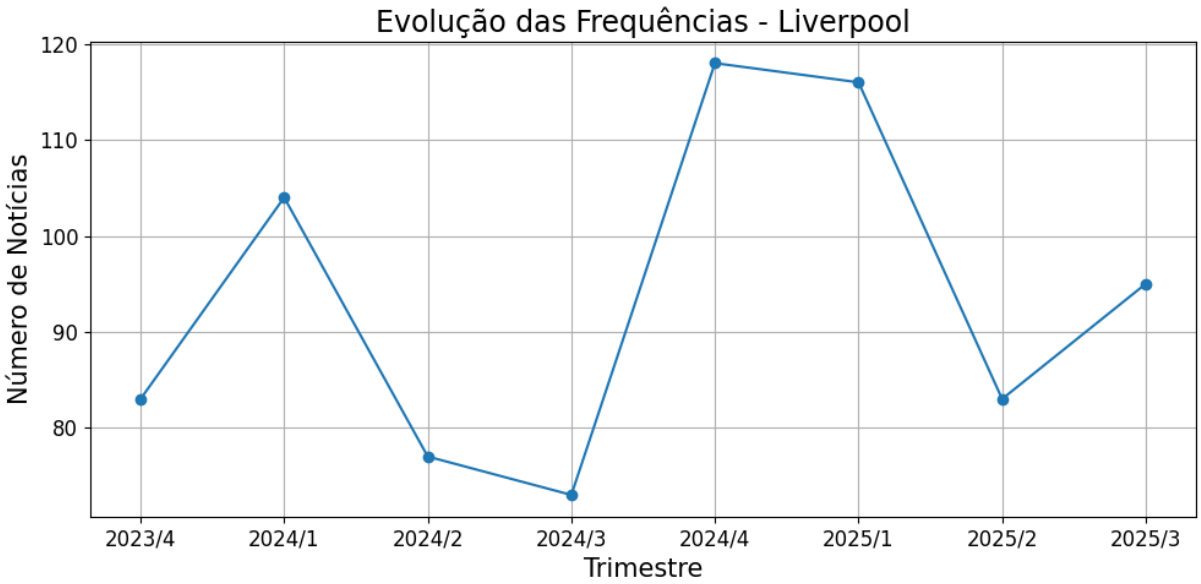

Figura 4.5: Número de notícias com a palavra "Liverpool".

por altos valores. Portanto, temos indicativos de que tais feitos levaram os holofotes da mídia a este clube.

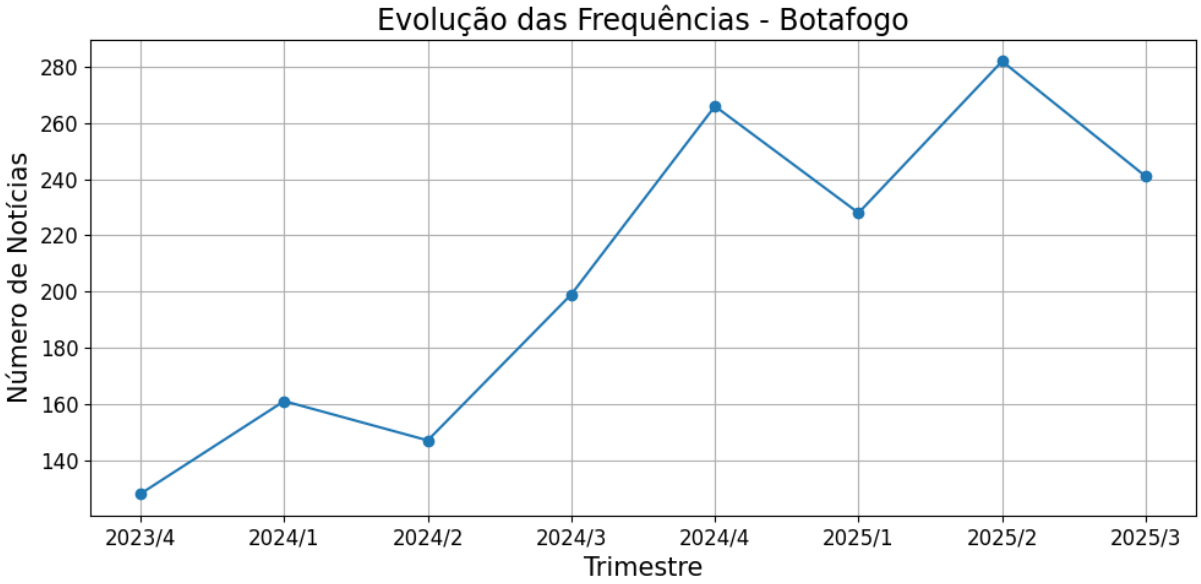

Figura 4.6: Número de notícias com a palavra "Botafogo".

Agora, analisando a evolução das frequências da palavra "Botafogo", classificado como ORG, na Figura 4.6 temos que conforme o tempo do estudo avança, o número de notícias citando a organização aumentou. Por este fato temos indicativos que os feitos que ocorreram durante o fim de 2024, títulos dos campeonatos Brasileiro e Libertadores, alavancaram o aporte midiático do clube, dobrando o número de citações do mesmo ao decorrer do período e mantendo-o neste patamar midiático durante o ano de 2025, com a

constância de notícias vista nesse ano.

4.2 Análises das Similaridades

As análises das frequências das palavras nos permite detalhar o número de entidades nos trimestres, porém sem o apoio de métricas para a justificativa para tal. Logo, uma forma de argumentar tais fatos é através das variações de similaridades que diferentes palavras tem com a palavra de referência durante os trimestres analisados. Ademais, com a aplicação das boas práticas de PLN e da vetorização, citada no Capítulo 3, conseguimos calcular tais similaridades e associar, por período, a possível causa do aumento ou diminuição das citações de uma palavra nas notícias.

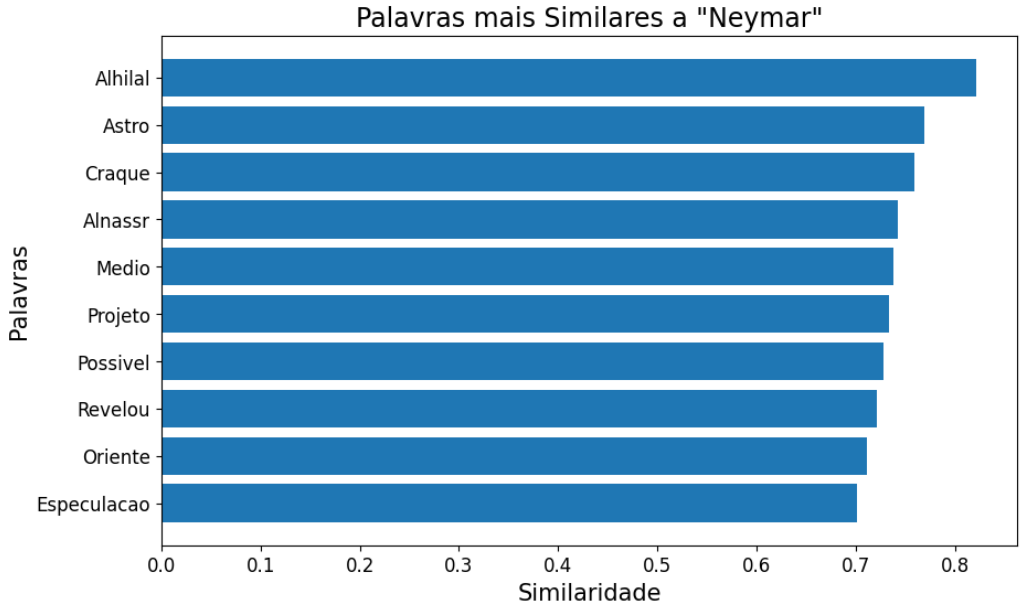

Figura 4.7: Top 10 palavras mais similares a "Neymar".

A Figura 4.7, dadas todas as notícias, representa as palavras que estão mais relacionadas a "Neymar" durante o período da pesquisa, por meio do cálculo da similaridade entre vetores. Por atuar quase todo o tempo em um clube do Oriente Médio, termos que são referentes a essa região ganham destaque na lista como "Alhilal" e "Alnassr", clubes da Liga da Arábia Saudita, e "Oriente" e "Medio", propriamente ditos. Além disso, "Astro" e "Craque" remetem a sua habilidade no futebol mundialmente conhecida no meio futebolístico que, apesar da diminuição do número de notícias totais sobre ele, ainda é amplamente conhecido por isso.

Ademais, ao analisar a palavra "Neymar", vimos anteriormente a variação de notí-

cias que essa palavra teve. Agora, com o cálculo das similaridades de algumas palavras selecionadas pelo autor, conseguimos ter alguns indicativos.

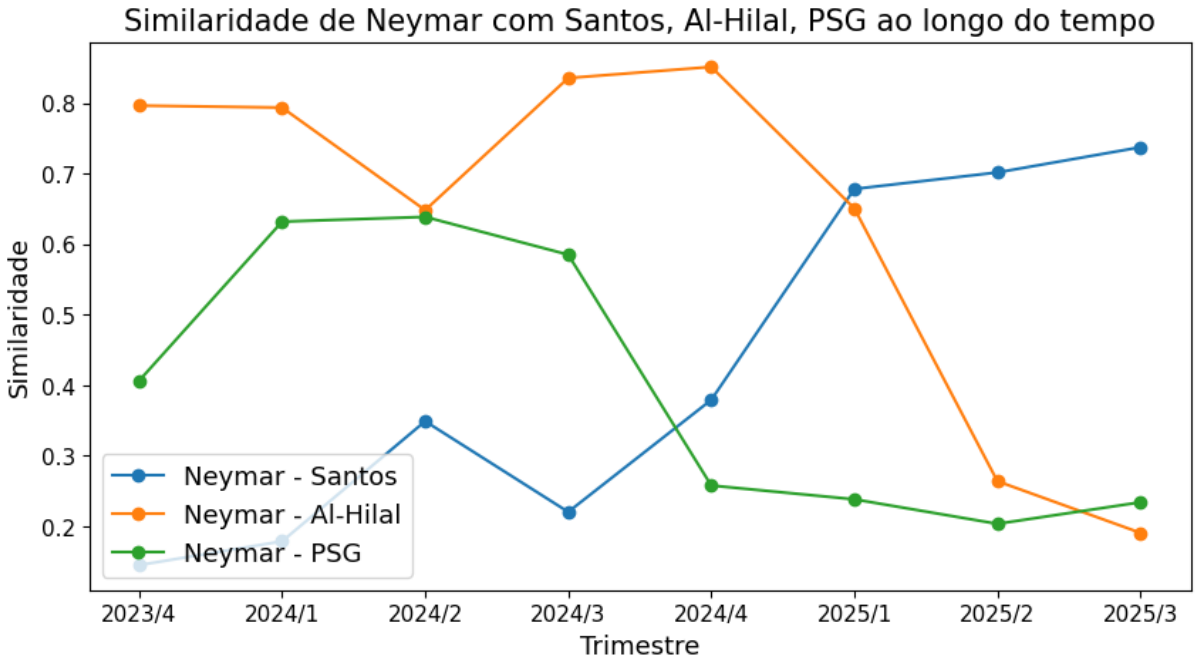

Figura 4.8: Similaridade de "Neymar" com clubes.

A Figura 4.8 fornece alguns indicativos sobre a palavra de referência e os últimos clubes em que o jogador atuou. A palavra "PSG" se refere a um clube de futebol que este atleta jogou antes do período analisado, por isso, vemos que esta palavra não está com a maior similaridade em qualquer trimestre, indicando que a relação do jogador com este clube diminuiu após sua saída. Em seguida, vemos que "Al-Hilal" se mantém como a palavra com maior similaridade até o fim do ano de 2024, afinal, era o clube em que Neymar estava atuando durante o período. Por fim, "Santos" apresentou um comportamento crescente nos valores de similaridade, onde o aumento do número a partir do fim de 2024 se deve aos rumores e oficialização do atleta como jogador do clube e, com isso, a relação entre Neymar e o clube se mantém alta durante o ano de 2025.

Outrossim, notamos que o comportamento da similaridade de Santos e Neymar é similar ao número de citações que "Neymar" teve no período de estudo (Figura 4.3). Logo, temos um indicativo de que o fato do atleta ter se transferido para o clube brasileiro, em particular, o seu primeiro clube como jogador profissional, fez com que as citações de seu nome nas notícias aumentassem significativamente, reforçando o seu grande aporte midiático citado na análise descritiva.

Por outro lado, a Figura 4.9 nos indica um fato interessante. Historicamente, este

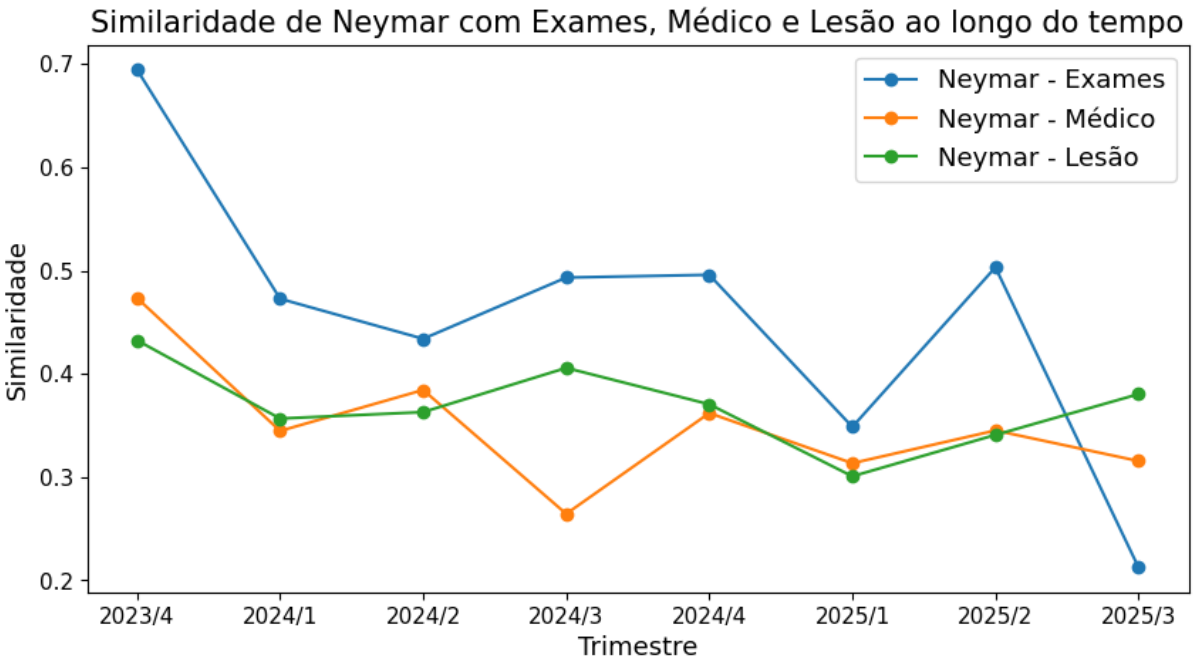

Figura 4.9: Similaridade de "Neymar" com termos que remetem à lesão.

profissional possuiu muitas lesões durante a sua carreira que acabaram por impedir sua atuação em diversas partidas, mas que geraram uma exposição extrema do mesmo nos anais futebolísticos. Porém, no período analisado vemos que este fato sofreu uma alteração; no primeiro trimestre do estudo, o jogador sofreu uma lesão e permaneceu praticamente o ano de 2024 todo sem atuar. Com este fato, temos o indicativo de que as maiores similaridades de expressões que remetem a lesões pertençam a este trimestre.

Ademais, também temos um indicativo de que o fato do número de notícias ter diminuído significativamente no ano de 2024 sobre esse jogador (Figura 4.3) pode estar associado ao fato do mesmo estar lesionado nesse período que, apesar de esperarmos uma alta similaridade entre estas palavras, resultou em um efeito contrário, pois a sua lesão fez com que seu aporte midiático diminuísse. Por fim, vemos que o pico da palavra "Exame" no Trimestre 2025/2, pode ter sido ocasionada por um machucado que o jogador sofreu e o impediu de atuar neste trimestre.

Agora, ao analisar a palavra "Messi", a Figura 4.10 reporta as principais palavras relacionadas ao jogador. Além do seu primeiro nome, "Lionel", vemos que alguns outros jogadores, "Benzema" e "Cristiano", que também tiveram seus auges em período semelhantes ao argentino, continuam altamente relacionados ao jogador. Igualmente conhecido por sua habilidade, as palavras "Embaixador" e "Astro" também remetem à sua fama no futebol. Por fim, o jogador possui alta similaridade com a cidade em que hoje está

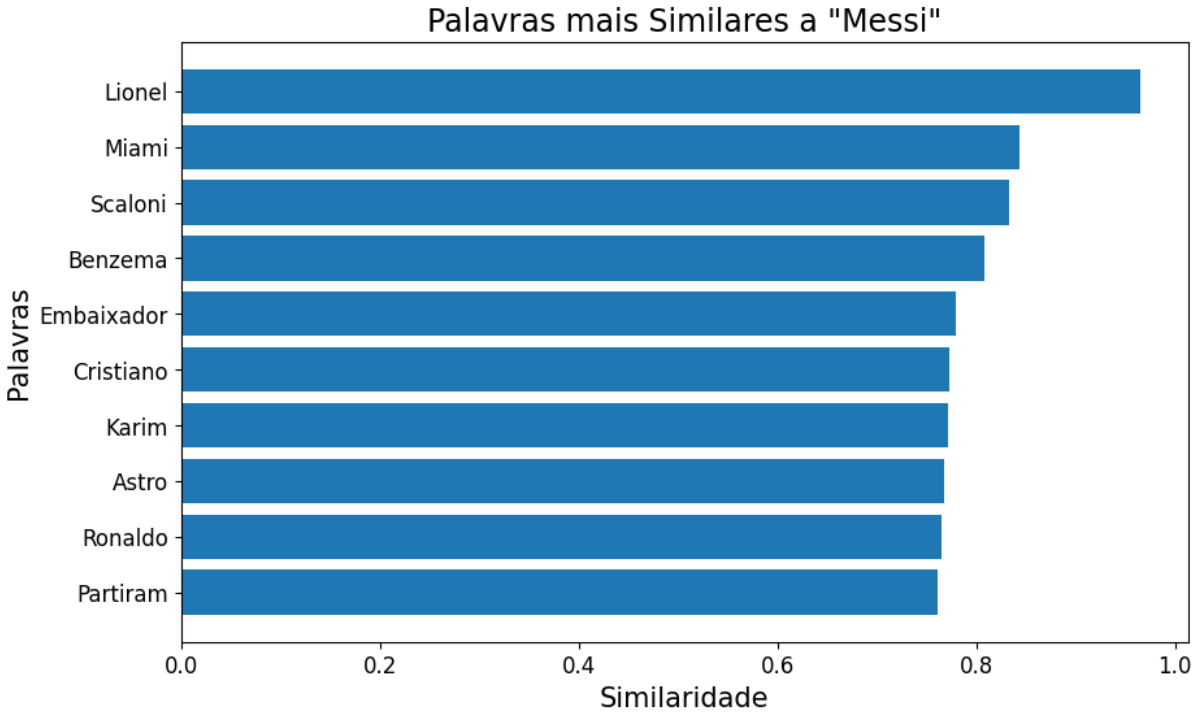

Figura 4.10: Top 10 palavras mais similares a "Messi".

o clube do mesmo, "Miami", indicando que as notícias estão relacionadas com ele e seu clube, Inter Miami.

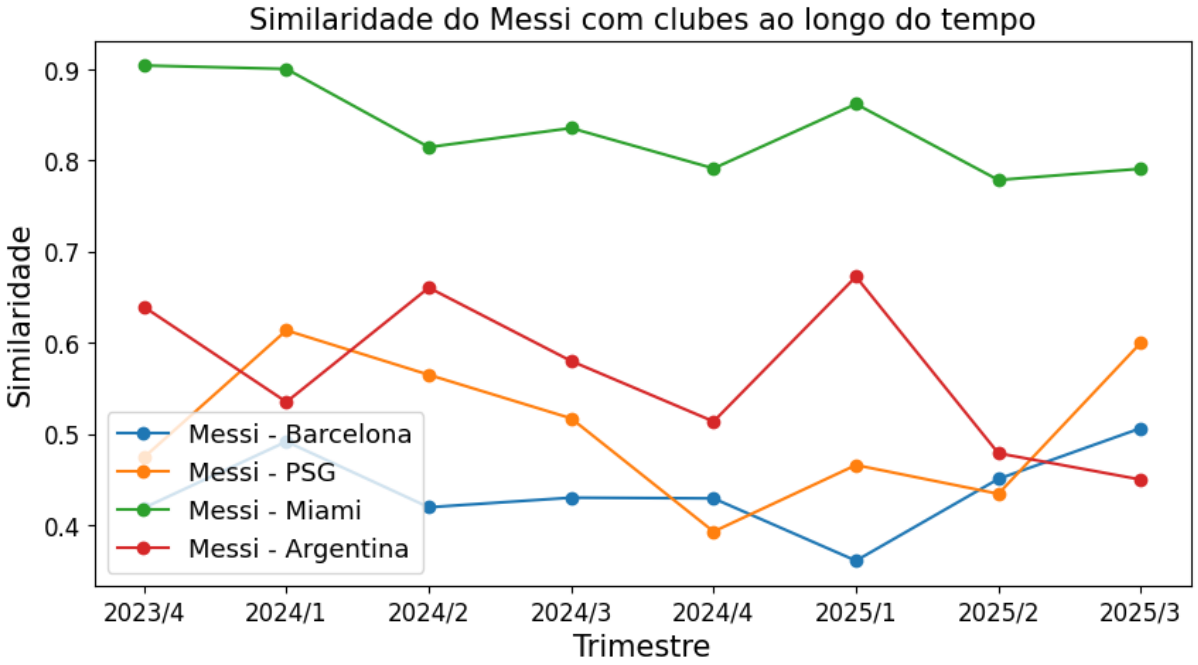

Figura 4.11: Similaridade de "Messi" com clubes.

Ademais, a Figura 4.11 nos indica as similaridades dos últimos clubes onde o jogador Lionel Messi atuou. "Barcelona" e "PSG" foram as últimas organizações em que o jogador

atuou antes de se transferir para os Estados Unidos, para o clube Inter Miami. Sobre estes, vemos que a relação com o jogador se tornou ínfima mesmo com sua recente saída do PSG e identificação com o clube espanhol. E, com a transferência ao clube americano, sua relação com a equipe se tornou o principal foco das notícias, com as similaridade entre eles sendo as maiores durante todo o período de estudo.

Logo, vemos que com sua saída do futebol Europeu, o número de notícias envolvendo seu nome caíram, justificado pela grande relação do jogador com o clube americano, que atua em um campeonato de pouca expressão no meio futebolístico. Além disso, observamos que o jogador sempre esteve a frente da Seleção de seu país, a Argentina, e, por meio das similaridades com essa palavra, vemos que as interações midiáticas não estão tão significativas para um jogador de tamanha representatividade em seu país. Portanto, este fato corrobora com a desvinculação do seu nome a Seleção com a iminente aposentadoria do jogador.

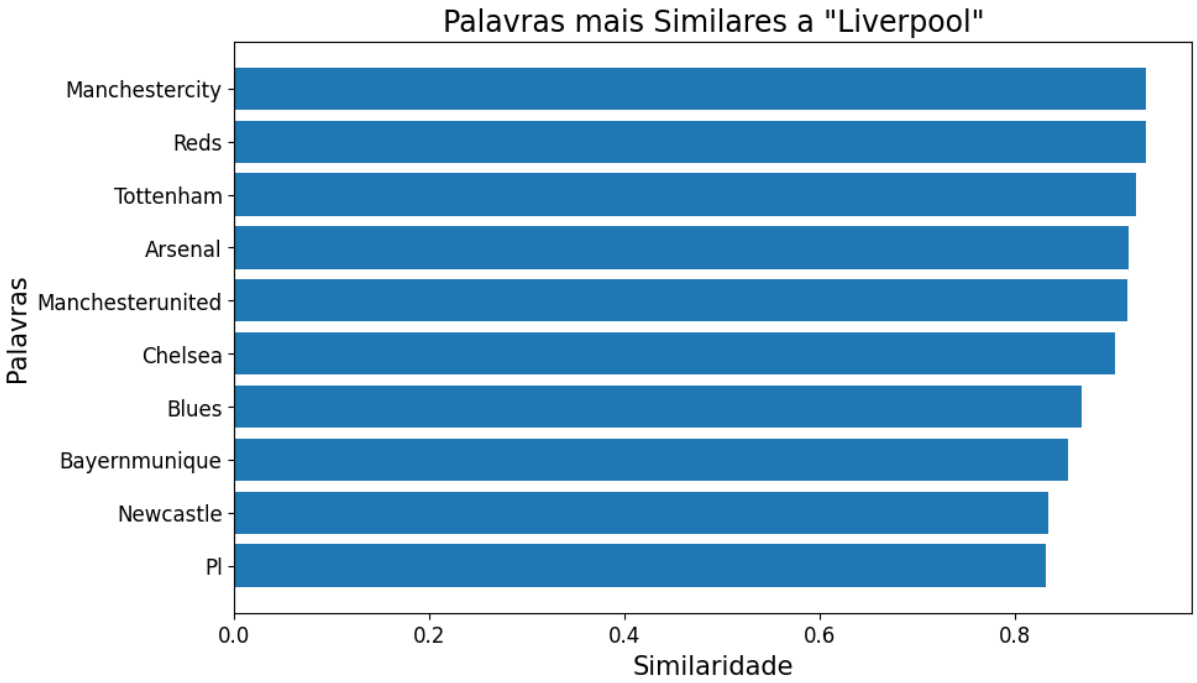

Figura 4.12: Top 10 palavras mais similares a "Liverpool".

Através da Figura 4.12 vemos que as principais palavras associadas ao clube Liverpool são de outros times europeus, em sua maioria ingleses. Por sua rivalidade recente com o Manchester City, temos indicativos de que esse fato prevaleceu na análise das notícias, pois a maior relação está com esse clube. Ademais, temos a alta similaridade com a palavra "Reds", pois é o nome como este clube inglês é apelidado no meio futebolístico.

Agora, com os indicativos obtidos sobre as transferências do clube, a Figura 4.13

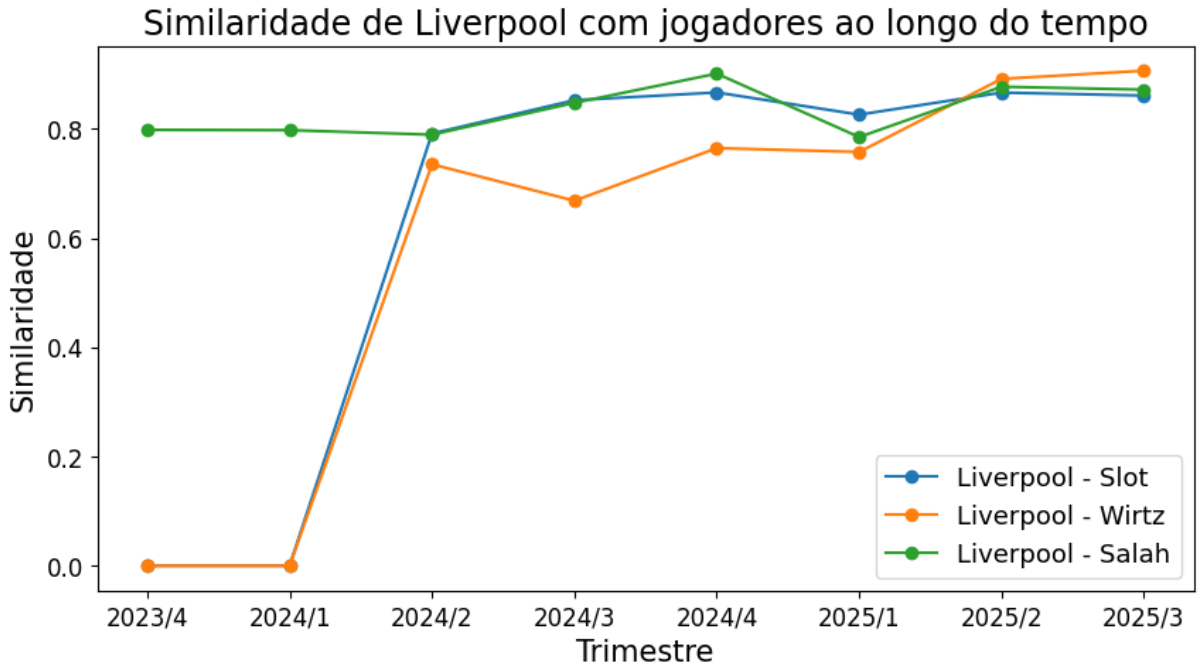

Figura 4.13: Similaridade de "Liverpool" com jogadores.

retrata três cenários. O primeiro é do jogador Salah que compõem o elenco do clube durante todo o período de análise e, portanto, possui uma alta identificação com o time durante este tempo. A segunda remete ao treinador Slot que foi contratado pelo clube em julho de 2024 e, como vemos no gráfico, nos primeiros semestres não há relação entre eles, pois as especulações de contratação se iniciaram no segundo semestre do mesmo ano. Por fim, o jogador Wirtz começou a ter seu nome relacionado ao Liverpool também no segundo semestre de 2024, porém concluiu sua transferência ao time apenas em 2025. Isso pode ser observado também, afinal a similaridade aumenta ainda mais nos dois últimos semestre de 2025 analisados.

Tais análises reforçam o indicativo de que, as movimentações de jogadores do Liverpool no mercado de transferências aumentaram o número de citações do clube nas notícias. Além disso, temos indicativos de que jogadores que já estavam no elenco ou mesmo o treinador, que foi contratado durante o período de análise, não contribuíram para a evolução do engajamento midiático significativamente, pois mesmo com Salah no elenco houveram variações no número de notícias e a chegada do novo treinador não alavancou o poder midiático do clube, que permaneceu baixo no trimestre da sua chegada.

Por fim, para a palavra "Botafogo", a Figura 4.14 demonstra que as principais palavras associadas ao clube, durante todo o período, são de outros clubes brasileiros que constantemente estão nas principais mídias. Outrossim, vemos que "Glorioso" também

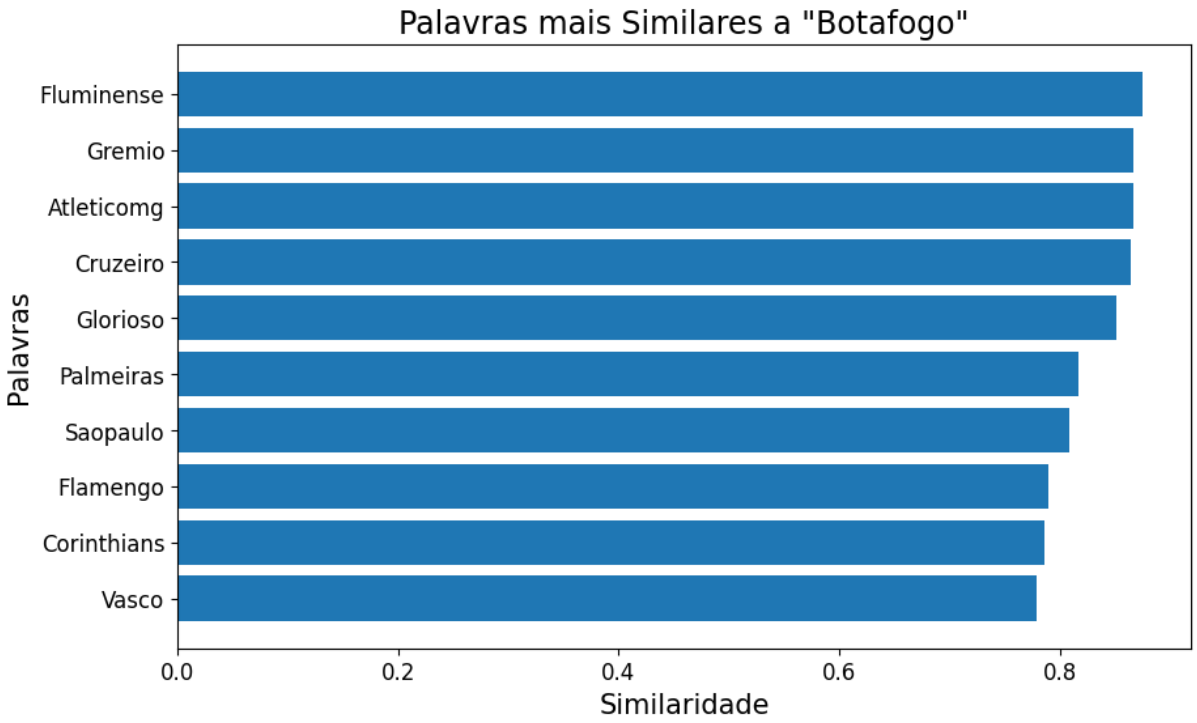

Figura 4.14: Top 10 palavras mais similares a "Botafogo".

está altamente correlacionado ao clube, afinal esse nome é o apelido do time no meio futebolístico.

Agora, a Figura 4.15 remete aos campeonatos disputados pelo Botafogo durante o período analisado. No ano de 2024, o Botafogo tinha uma performance superior aos demais clubes, liderando o Campeonato Brasileiro e avançando as fases da Copa Libertadores. Por esses motivos as similaridade entre estas palavras permaneceram altas durante o ano e, como estes campeonatos se findam no último trimestre do ano e o time ganhou ambos campeonatos, a palavra "Campeão" foi muito associado ao Botafogo. Além disso, vemos que este período vitorioso não continuou no ano seguinte, afinal a similaridade com esta palavra decaiu significativamente. Outrossim, a palavra "Mundial" tem um aumento significativo na relação com o clube no semestre 2024/4, período em que o time estava disputando o campeonato intercontinental Mundial de Clubes.

Logo, temos indicativos de que tais palavras explicam o comportamento da evolução de notícias que o clube teve neste período. A evolução observada no final de 2024 está em concordância com o período de conclusão dos campeonatos e conquista dos títulos do Botafogo que, consequentemente, impactam no número de notícias que o clube é citado.

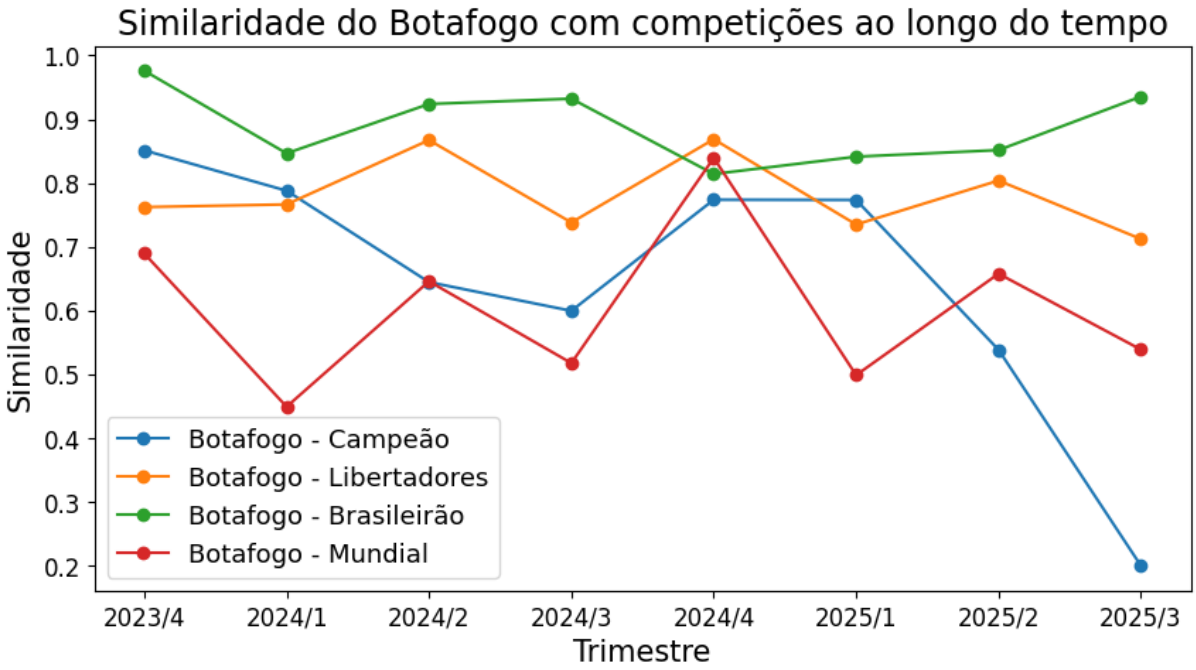

Figura 4.15: Similaridade de "Botafogo" com campeonatos.

4.3 Reduções de Dimensionalidade

As análises anteriores resultaram em indicativos locais, específicos para cada termo analisado. As reduções de dimensionalidade a serem apresentadas nesta seção buscarão representações globais sobre algumas das principais palavras contidas nos textos, revelando proximidades das mesmas e, consequentemente, suas relações. Serão abordadas as técnicas de Componentes Principais (PCA) e de Vizinhos Estocásticos (t-SNE).

Iniciando as análises, a Figura 4.16 representa as entidades que são classificadas como PER. Notamos que alguns jogadores que atuam dentro e fora do futebol brasileiro podem ser separados pelo Componente Principal 1. Logo, as palavras com valores negativos desse componente tendem a serem jogadores de clubes europeus e, para os positivos, jogadores de times do Campeonato Brasileiro. Isto nos indica que nas notícias, estes jogadores são citados em contextos diferentes.

A Figura 4.17 representa o PCA aplicado nas entidades classificadas como LOC. Nele, vemos que as cidades e países tendem a ser divididos por meio do Componente Principal 1, onde as com valores positivos, como Brasil, Bolívia e Santos, pertencem à América do Sul e com valores negativos estão concentradas na Europa, como Barcelona, Inglaterra e Paris. Isto nos indica comportamentos distintos entre os clubes e jogadores nos textos publicados pelos jornalistas.

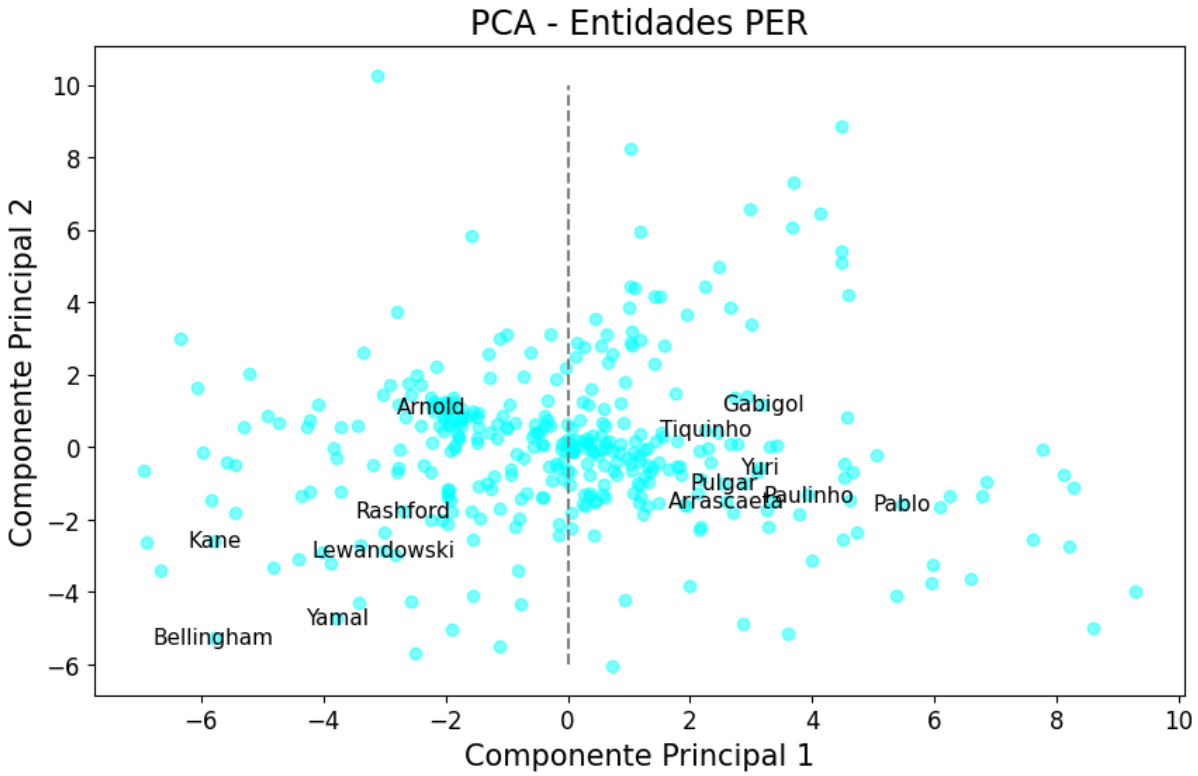

Figura 4.16: PCA para entidades PER.

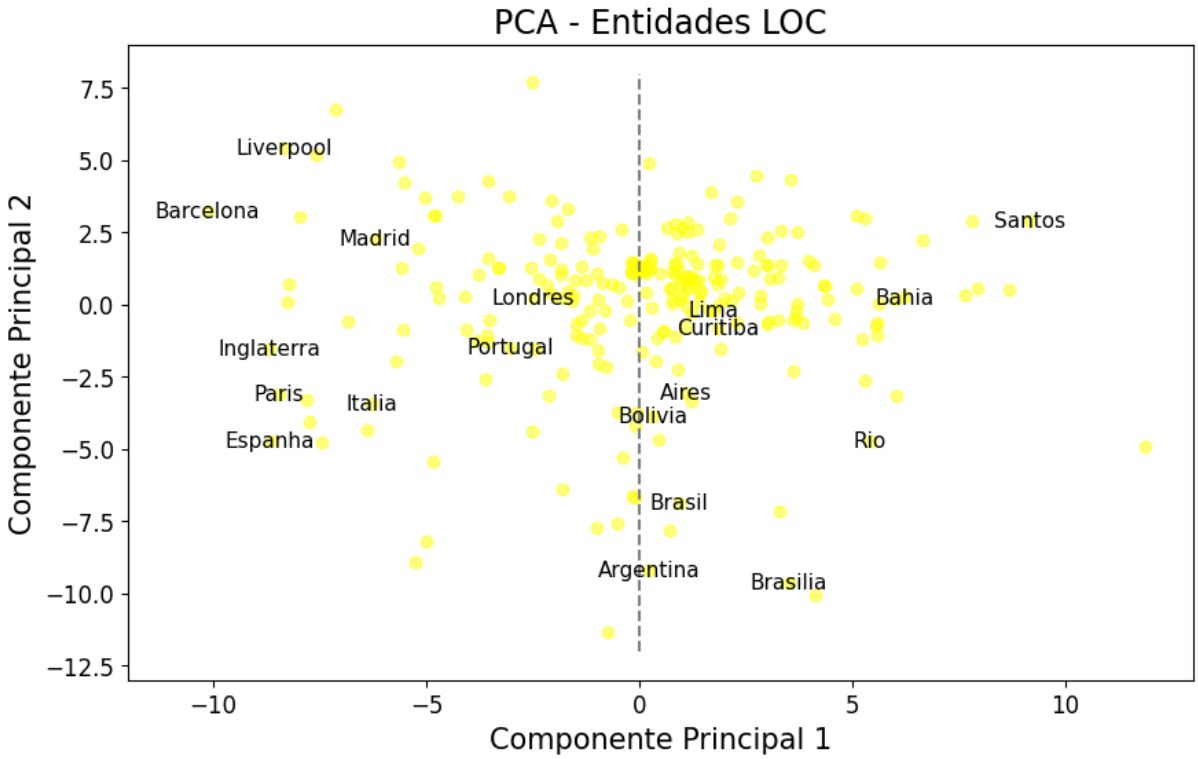

Figura 4.17: PCA para entidades LOC.

Da mesma forma como observado anteriormente, a Figura 4.18 revela a distribuição dos pontos da categoria ORG e, por meio do Componente Principal 1, notamos uma

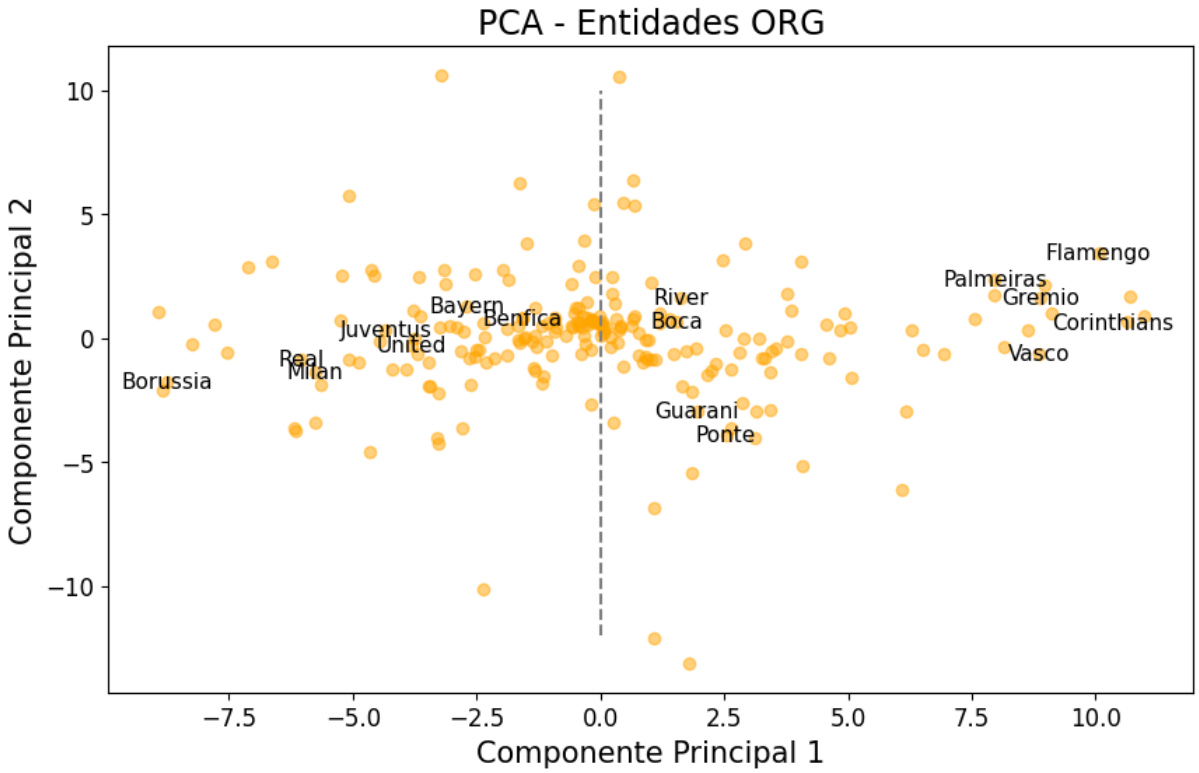

Figura 4.18: PCA para entidades ORG.

tendência de segmentação entre os clubes de futebol sediados na Europa e na América do Sul. Os times sul-americanos se concentram nos valores negativos do componente, enquanto os europeus estão com números positivos para o mesmo.

Além disso, podemos observar pontos que estão aglomerados no gráfico e que também tem alta relação no mundo do futebol. Primeiramente, os times brasileiros estão com seus respectivos pontos próximos uns aos outros, nos indicando que seus comportamentos nas notícias esportivas são semelhantes. Este fato é razoável, pois são clubes que corriqueiramente são citados no Mercado da Bola e por resultados de seus jogos, sendo derrotas ou vitórias, além da grande força midiática que possuem, por atuarem no principal campeonato do Brasil. Outra relação se dá pela proximidade de Guarani e "Ponte", onde "Ponte" remete ao clube Ponte Preta. As similaridades desses clubes são altas no futebol, pois ambos são sediados na cidade de Campinas e formam uma das maiores rivalidades do Brasil. Por fim, as palavras "Boca" e "River" também remetem a dois clubes conhecidos por sua rivalidade. Boca Juniors e River Plate são dois times argentinos que assumem comportamentos similares por atuarem no mesmo país, sendo duas das principais potências da Liga Argentina e antagonistas em seus duelos.

Em seguida, abordaremos uma outra técnica de redução de dimensionalidade, o t-

SNE, para encontrar possíveis novos padrões que não foram identificados na aplicação da técnica anterior. Para isso, o processo contém dois hiper-parâmetros principais, o número de vizinhos a serem considerados na estimação das probabilidades e o número de iterações da função de otimização. Ambos foram escolhidos neste trabalho de maneira a gerar resultados interpretáveis para as análises e, por meio de testes feitos, valores pequenos para os mesmos originaram melhores soluções. Juntamente com isto, o método de clusterização K-médias será aplicado para a visualização dos agrupamentos entre as palavras.

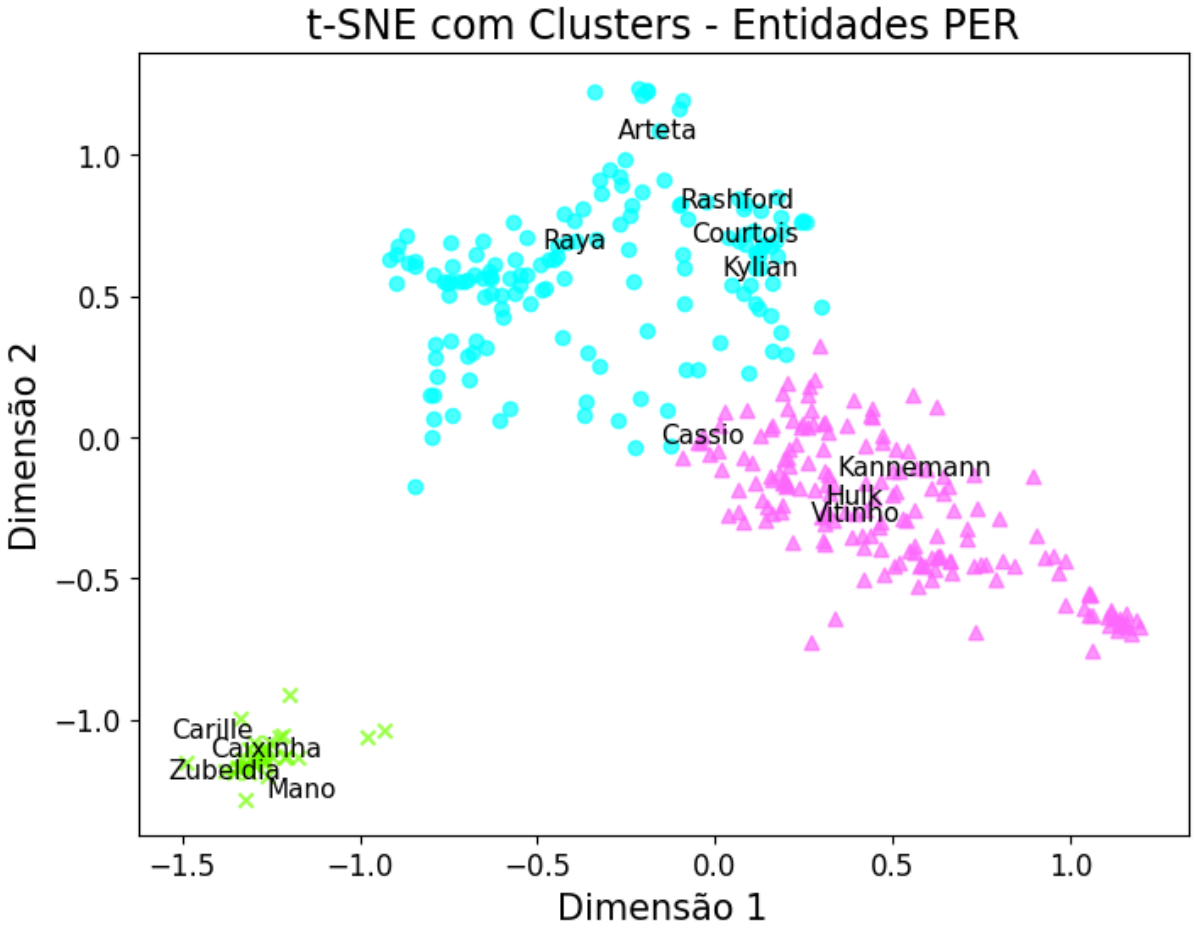

Figura 4.19: Clusterização do t-SNE para entidades PER.

A Figura 4.19 representa as entidades PER e, como podemos analisar, a clusterização segmentou as palavras tendendo a seguir alguns critérios. Ao cluster em tom esverdeado, observamos a tendência de concentração de técnicos que atuam no futebol brasileiro. Já para o cluster rosado, há a tendência de conter jogadores que atuam no futebol brasileiro e, portanto, separando os técnicos dos jogadores do futebol nacional, característica que não foi possível observar nas análises anteriores, mas que também indica diferentes com-

portamentos entre estes profissionais. E, por fim, o último cluster tende a conter tanto jogadores quanto treinadores que atuam no futebol europeu.

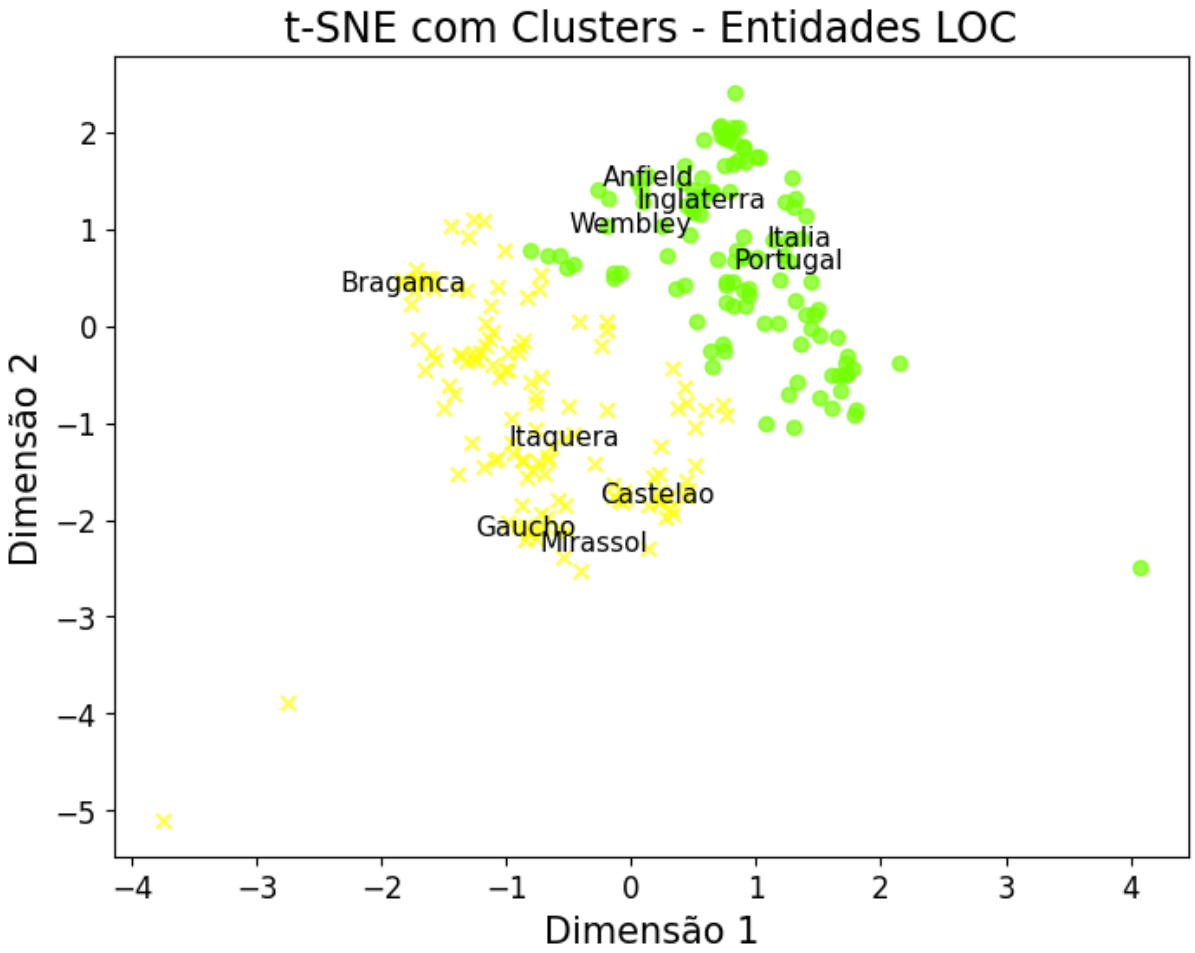

Figura 4.20: Clusterização do t-SNE para entidades LOC.

Na Figura 4.20, que retrata as entidades classificadas como LOC, podemos ver claramente a tendência das localidades. Na clusterização, o grupo em tom amarelado tende a conter localidades (cidades, estádios, entre outros) que pertencem ao território brasileiro. Por outro lado, o outro agrupamento tende a possui locais (cidades, estádios, países) que estão no continente europeu. Isto traz, novamente, indicativos de que estas localidades destas regiões tem comportamentos distintos nas notícias.

Por fim, na Figura 4.21 temos a representação das entidades classificadas como ORG, sendo em sua maioria, clubes de futebol. Nela, também observamos a formação de 2 cluster. O agrupamento em tons alaranjados tende a conter clubes de futebol que possuem sede em países europeus. Por outro lado, o grupo em tom esverdeado tende a conter times que são localizados na América do Sul. Esta análise reforça o indicativo obtido anteriormente, de que as notícias sobre clubes, jogadores e técnicos que atuam na Europa

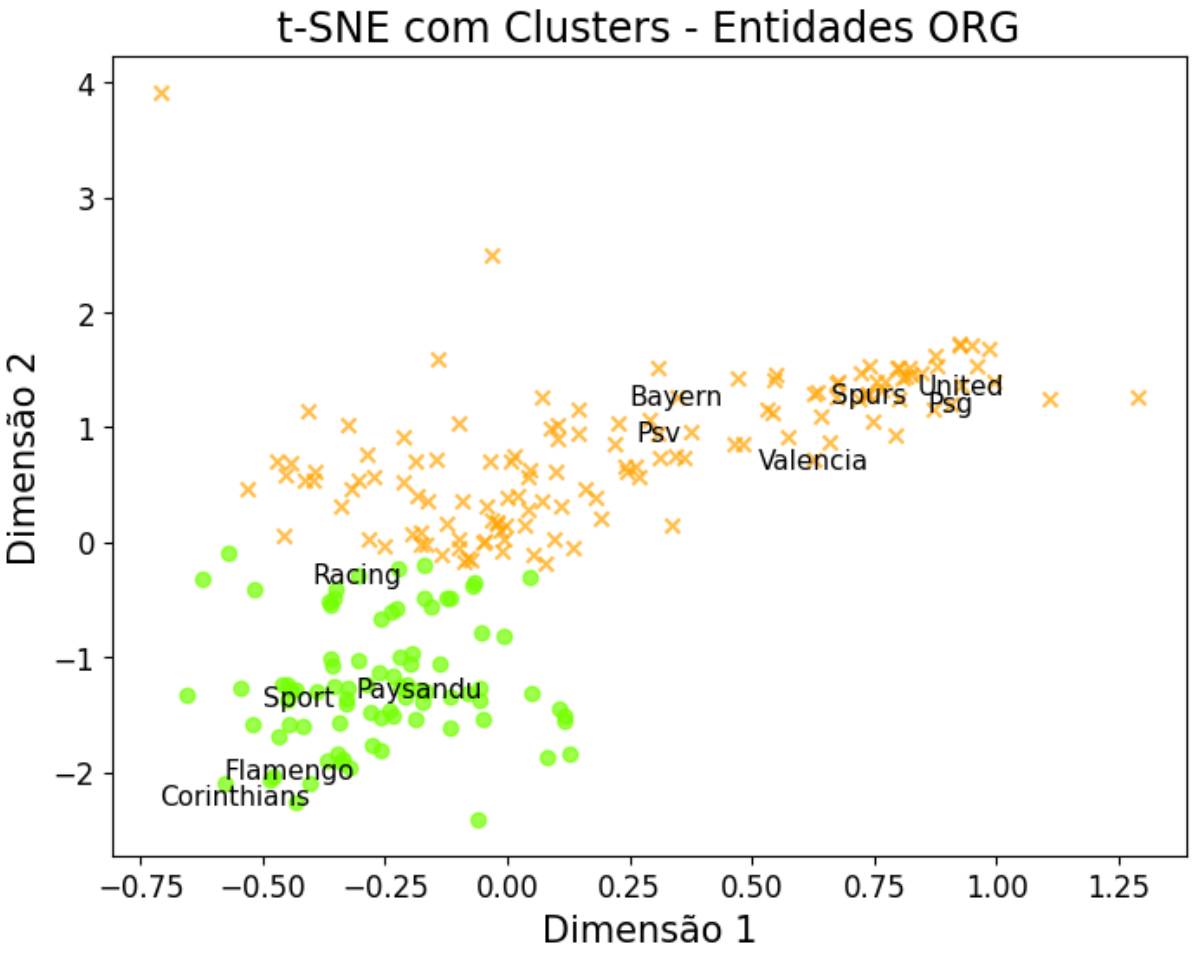

Figura 4.21: Clusterização do t-SNE para entidades ORG.

tendem a ser diferentes das reportagens publicadas sobre os profissionais que atuam no futebol brasileiro ou na América do Sul, de modo geral.

4.4 Equações matemáticas

No Capítulo 3 vimos que por meio da vetorização é possível realizar equações matemáticas entre os vetores das palavras encontrados. Com isso, podemos encontrar similaridades entre palavras que são irrelevantes, a priori, mas quando retiramos o efeito de uma palavra descobrimos tais relações. A primeira equação a ser analisada será,

$$Mbappe - Europa + Brasil,$$

e, desta maneira, buscando encontrar algum jogador semelhante a profissional Mbappé. Logo retirando o efeito do continente do seu país, Europa, e adicionando uma palavra de interesse, Brasil encontramos algumas palavras que possuem maior similaridade com

esta operação e nos remetem ao jogador brasileiro Vinícius Júnior (Figura 4.22). Afinal, podemos interpretar que quatro termos tem grande similaridade, "Jr", "Vini", "Vinicius" e "Junior".

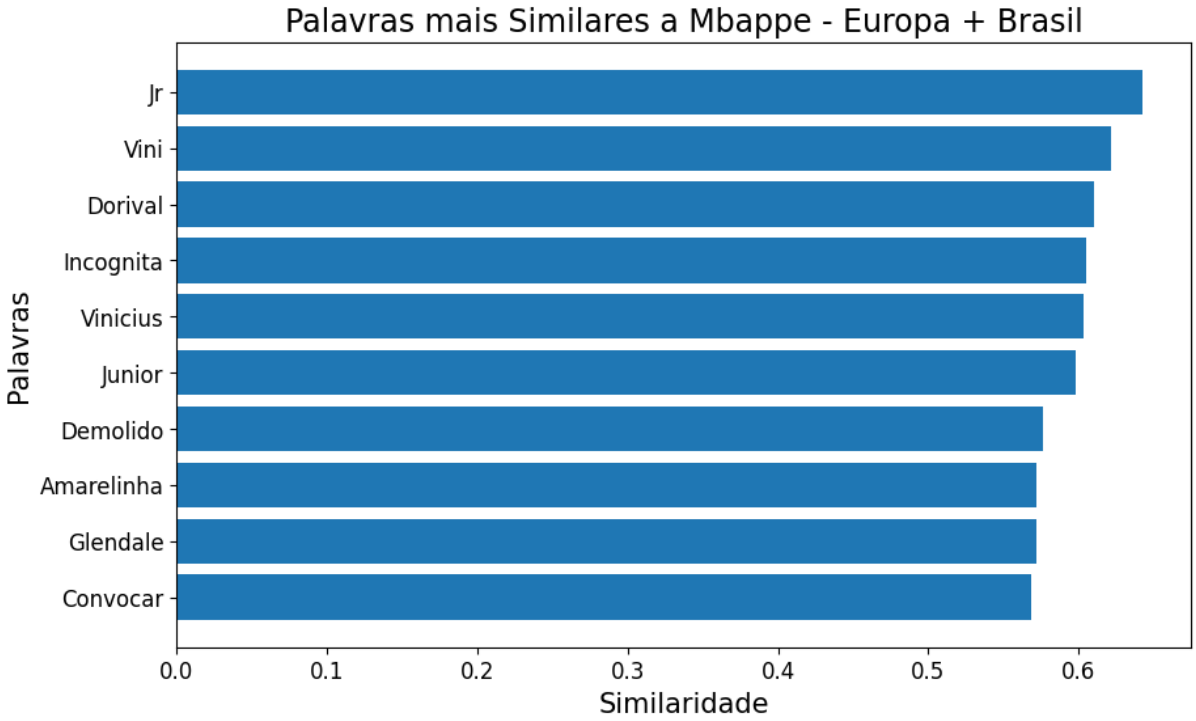

Figura 4.22: Equação para a palavra "Mbappe".

Agora, ao centralizar a análise sobre o treinador brasileiro Dorival, a seguinte equação dos vetores das palavras foi feita,

$$Dorival - Brasil + Europa,$$

com isso, buscando retirar o fato do treinador ser brasileiro e querendo encontrar os termos mais similares ao adicionarmos a palavra "Europa".

Portanto, pela Figura 4.23, vemos que a grande parte dos nomes citados como os mais similares são de treinadores de nacionalidades europeias. Logo, a vetorização obteve uma boa acurácia em encontrar relação entre técnicos, mesmo que possuam nacionalidades diferentes e, muita vezes, não são citados juntos nas mesmas notícias. Ademais, observamos que a equação preservou a similaridade da palavra de referência, "Dorival", com os dois nomes dos quais os treinadores são conhecidos, "Jurgen Klopp", "Pep Guardiola", "Hansi Flick" e "Carlo Ancelotti", que apesar do nome "Ancelotti" não estar no gráfico, aparece como a 14ª palavra mais similar à equação.

Por fim, vamos a analisar a equação sobre a equipe do Mirassol, clube que tem crescido

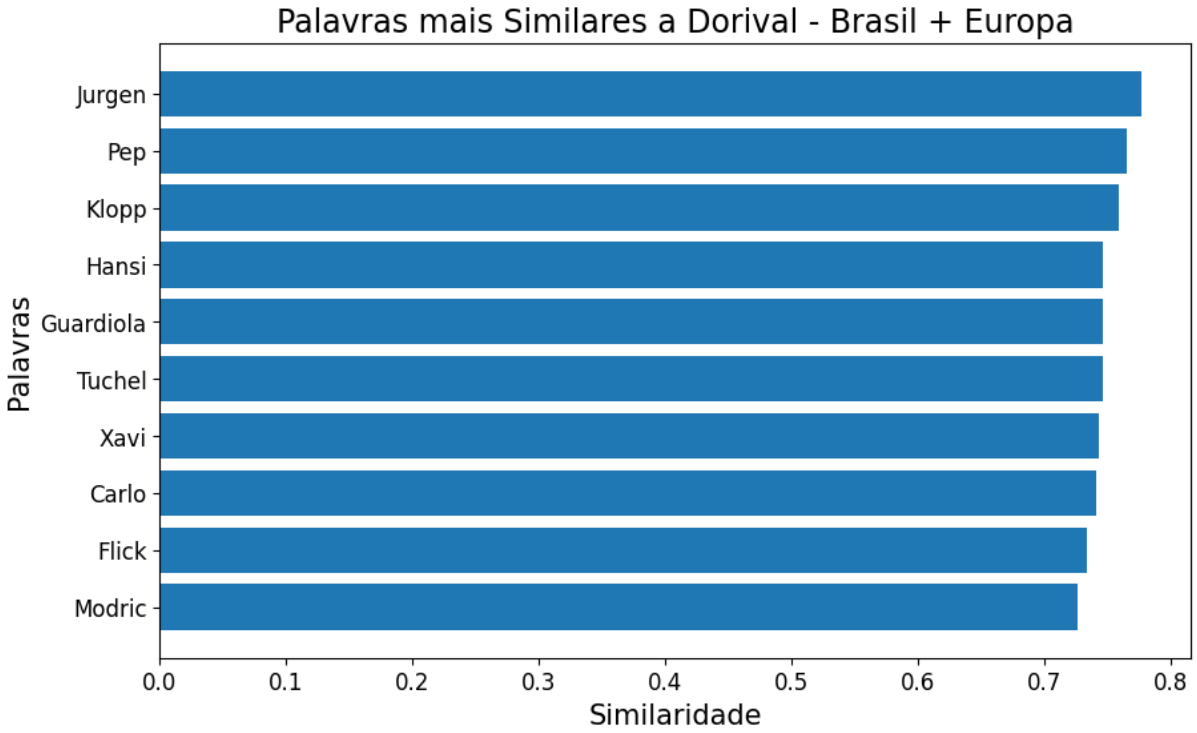

Figura 4.23: Equação para a palavra "Dorival".

de rendimento e de visibilidade no cenário nacional, através da equação,

$$Mirassol - Brasil + Inglaterra.$$

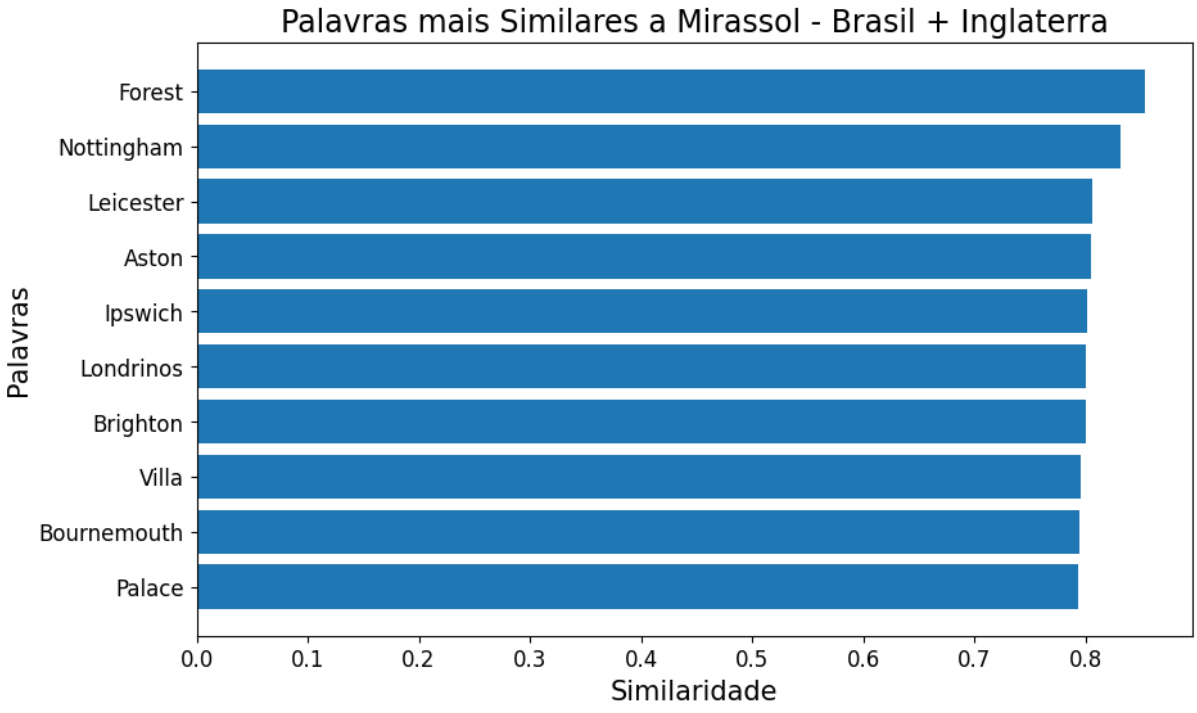

Figura 4.24: Equação para a palavra "Mirassol".

A Figura 4.24 mostra que o clube Mirassol, ao realizar o cálculo, possui uma grande relação com a equipe inglesa do Nottingham Forest, sendo os dois termos com maiores similaridades analisados. É interessante observar que, durante uma boa parte do período de estudo, o clube da Inglaterra teve um bom desempenho no campeonato local, mesmo sendo considerado um clube de menor expressão. Por outro lado, a equipe brasileira teve resultados muito bons, tanto no ano de 2024, conquistando a classificação para o campeonato da primeira divisão, quanto no ano de 2025, onde estava até a última data analisada, tendo um resultado surpreendente para um clube debutante do Campeonato Brasileiro da Série A. Logo, vemos que nos últimos anos, os desempenhos destes clubes são equiparáveis em seus respectivos campeonatos nacionais e como as notícias retratam este fato, pois ao analisar os termos que remetem a estes clubes eles possuem alta similaridade, mas que era ocultada por efeitos de outras palavras.

Capítulo 5

Conclusão

O Processamento de Linguagem Natural mostra-se uma ferramenta eficaz para a análise do principal meio de comunicação no mundo futebolístico na atualidade, em que a internet possibilita a disseminação de notícias de forma eficiente. A própria organização e estruturação dos dados se tornam mais claras graças à agilidade na coleta das informações, potencializada pelo uso do *Web Scraping*, que permite uma extração assertiva e acessível a qualquer usuário disposto a aplicá-lo. Com as palavras e técnicas adequadas, os *insights* obtidos sobre os acontecimentos no futebol, somados ao conhecimento do leitor, resultam em representações fiéis da realidade dentro dos textos analisados. Nesse sentido, as evoluções das citações de jogadores, técnicos e clubes observadas ao longo deste trabalho alinham-se aos fatos ocorridos no período estudado, evidenciando que as notícias, a priori, avançam naturalmente conforme os próprios acontecimentos vão se desenvolvendo.

As mudanças de *status* dos jogadores Neymar e Messi traduzem o que vemos na prática. O primeiro jogador sendo cada vez mais esquecido no clube em que estava, Al-Hilal, e com sua transferência para o Santos, retorna aos holofotes da mídia. O segundo por sua vez, indicando que o auge da sua vida profissional já passou e o aporte midiático sobre ele já não é mais o mesmo, em uma liga de futebol alternativa e com a aposentadoria dos gramados se aproximando. Já para os clubes, vemos como os rumores de transferências e títulos alavancam os times na mídia e aumentam a exposição dos mesmos.

Por fim, de modo geral, clubes e jogadores que atuam no mesmo país ou no mesmo continente tendem a estar nas notícias de maneira similar. Esta segmentação foi bem observada nas análises, onde Europa e América do Sul se mostraram distintas durante grande parte das análises, mostrando que, as notícias desta plataforma separam, de maneira clara, textos sobre os estrangeiros e textos sobre o futebol nacional.

Referências Bibliográficas

90min (2025). 90min. https://www.90min.com/pt-br/noticias-futebol?page=160. Acessado em 30 de junho de 2025.

Aarsen, T., Nothman, J., Bird, S., Dimitradis, A., Sepler, D., Milajevs, D., Bond, F., Kurenkov, I. e Tan, L. (2024). Nltk. https://www.nltk.org/. Acessado em 23 de junho de 2025.

Afonso, A. R. e Duque, C. G. (2020). Mineração de textos aplicada a postagens do twitter sobre coronavírus: uma análise na linha do tempo. *Liinc*, **16**(2).

Bengio, Y., Ducharme, R., Vincent, P. e Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.

Cordasco, I. S., Benfield, C. e Prewitt, N. (2024). requests. https://pypi.org/project/requests/. Acessado em 04 de maio de 2025.

da Silva Ferreira, C. e Sampaio, D. A. (2018). As práticas sociais de linguagem e o Twitter: um estudo em dois jornais de Salvador-BA. *Revista Desempenho*, **2**(20), 1–8.

Grishman, R. e Sundheim, B. (1996). Message understanding conference-6: A brief history. *Proceedings of the 16th Conference on Computational Linguistics*, **1**, 466–471.

Haykin, S. (1994). Neural networks: A comprehensive foundation. *Prentice Hall*.

Honnibal, M. (2024). spacy: Industrial-strength natural language processing in python. https://spacy.io/ [Acesso em: 12 fev. 2025].

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**(6), 417–441.

Hutchins, W. J. (2004). The Georgetown-IBM experiment demonstrated in January 1954. *Springer Verlag*.

Izbicki, R. e dos Santos, T. M. (2022). *Aprendizado de Máquina: uma abordagem estatística*. UICLAP.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of ECML*.

Kosala, R. e Blockeel, H. (2000). *Web Mining Research: A Survey*, volume 2. ACM SIGKDD Explorations Newsletter.

Kullback, S. e Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

LeCun, Y., Bottou, L., Bengio, Y. e Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

Manning, C., Raghavan, P. e Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martin, J. H. e Jurafsky, D. (2025). *Speech and Language Processing*. Prentice Hall, 3st edition.

Mikolov, T., Chen, K., Corrado, G. e Dean, J. (2013). Efficient estimation of word representations in vector space. *[S.l.]: arXiv:1301.3781*.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(11), 559–572.

Rehurek, R. (2024). gensim. https://pypi.org/project/gensim/. Acessado em 07 de julho de 2025.

Richardson, L. (2025). bs4. https://pypi.org/project/beautifulsoup4/. Acessado em 30 de junho de 2025.

Salton, G., Wong, A. e Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

Schank, R. e Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum Associates.

Sudhakar, M. e Kaliyamurthie, K. (2024). Detection of fake news from social media using support vector machine learning algorithms. *Measurement: Sensors*, **32**, art. 101028.

van der Maaten, L. e Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, **9**, 2579–2605.