



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Alisson Fabra da Silva

**Caracterização e análise de formação de comunidades no contexto da Copa
do Mundo de 2022**

Florianópolis
2023

Alisson Fabra da Silva

Caracterização e análise de formação de comunidades no contexto da Copa do Mundo de 2022

Trabalho de Conclusão de Curso do Curso de Graduação em Ciências da Computação do Centro Tecnológico da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Ciências da Computação.

Orientador: Prof^ª. Carina Friedrich Dorneles, Dra.

Coorientador: Prof^ª. Ana Paula Couto da Silva, Dra. (UFMG)

Florianópolis

2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Silva, Alisson Fabra da
Caracterização e análise de formação de comunidades no
contexto da Copa do Mundo de 2022 / Alisson Fabra da Silva
; orientadora, Carina Friedrich Dorneles, coorientadora,
Ana Paula Couto da Silva, 2023.
128 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro Tecnológico,
Graduação em Ciências da Computação, Florianópolis, 2023.

Inclui referências.

1. Ciências da Computação. 2. Análise de dados. 3.
Detecção de comunidades. 4. Twitter. 5. Copa do Mundo. I.
Dorneles, Carina Friedrich . II. Silva, Ana Paula Couto
da. III. Universidade Federal de Santa Catarina. Graduação
em Ciências da Computação. IV. Título.

Alisson Fabra da Silva

Caracterização e análise de formação de comunidades no contexto da Copa do Mundo de 2022

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de “Bacharel em Ciências da Computação” e aprovado em sua forma final pelo Curso de Graduação em Ciências da Computação.

Florianópolis, 7 de dezembro de 2023.

Prof^a. Lúcia Helena Martins Pacheco, Dra.
Coordenadora do Curso

Banca Examinadora:

Prof^a. Carina Friedrich Dorneles, Dra.
Orientadora

Prof^a. Ana Paula Couto da Silva, Dra.
Coorientadora
UFMG

Prof^a. Mirella Moura Moro, Dra.
Avaliadora
UFMG

Prof^a. Michele Amaral Brandão, Dra.
Avaliadora
IFMG

AGRADECIMENTOS

Gostaria de expressar minha gratidão a todos que contribuíram e estiveram ao meu lado durante minha jornada acadêmica e na conclusão deste trabalho. Obrigado por me apoiarem nessa trajetória, cada um de vocês desempenhou um papel muito importante para mim.

Primeiramente, sou imensamente grato à minha orientadora Carina e à minha coorientadora Ana, cuja orientação, apoio e valiosos insights foram fundamentais para o desenvolvimento deste trabalho.

As acadêmicas Larissa Malagoli e Beatriz Paiva pelo apoio dado no desenvolvimento deste trabalho e pela disponibilização do conjunto de dados utilizado.

Aos meus pais, Alcionei e Maria Imaculada, por todas as oportunidades, investimentos e apoios que foram necessários para tornar possível toda essa jornada.

Ao meu irmão Nick, por todo o apoio, parceria, incentivo, orientação e disponibilidade sempre que precisei. Sua presença e encorajamento foram fundamentais para superar os desafios que encontrei até hoje.

A minha namorada, Elis, por todo o amor, carinho, apoio, compreensão e incentivo que me proporcionou sempre. Você é meu porto seguro. Sou profundamente grato por ter alguém tão especial como você ao meu lado.

Aos meus amigos, especialmente Diego e Duds, por todas as conversas, brincadeiras e apoios que me deram durante esse período. Em cada desafio, vocês estiveram ao meu lado, trazendo leveza e felicidade.

A minha gatinha Nina, por todo o companheirismo que me deu ao longo desses anos. Você enche meus dias de alegria.

Ao Laboratório Bridge, que me proporcionou um crescimento profissional muito grande, além de ter me apresentado muitos amigos. Muito obrigado por todas as oportunidades.

A Universidade Federal de Santa Catarina e aos professores que foram fundamentais na minha formação, agradeço pela rica troca de conhecimento e pelas oportunidades valiosas concedidas.

RESUMO

Através das redes sociais, muitos ganham popularidade ao expressar suas opiniões e atrair seguidores, formando uma comunidade em torno dessas ideias. O propósito deste trabalho é a utilização de um algoritmo de detecção de comunidades em um conjunto de dados provenientes de opiniões de pessoas de uma rede social, que abordam o mesmo tema. O objetivo é analisar o comportamento dos internautas para compreender o que está sendo discutido em determinados grupos (ou comunidades). Através da análise dos resultados obtidos após a aplicação dos algoritmos, seria possível entender o comportamento dos membros da rede em situações do mundo real. As técnicas de detecção de comunidades são úteis para identificar pessoas com interesses semelhantes e mantê-las conectadas, além de permitir a extração de grupos com características similares. O conjunto de dados utilizados foi coletado do Twitter no período pré, durante e pós Copa do Mundo de 2022. No desenvolvimento dessa análise são utilizados grafos, de modo a usufruir de algumas propriedades, como as comunidades e a identificação de indivíduos ativos e influentes através das arestas que chegam e saem de um vértice. Para atingir o objetivo, inicialmente é feita a caracterização da amostra de dados coletados com o objetivo de compreender os dados e obter uma sumarização de suas principais características. O resultado ao final do trabalho é a identificação e análise das comunidades formadas nesse contexto.

Palavras-chave: Análise de dados. Detecção de comunidades. Twitter. Copa do Mundo.

ABSTRACT

Through social media, many gain popularity by expressing their opinions and attracting followers, forming a community around these ideas. The purpose of this work is to utilize a community detection algorithm on a dataset derived from opinions of individuals on a social network discussing the same topic. The objective is to analyze the behavior of internet users to understand what is being discussed in specific groups (or communities). By analyzing the results obtained after applying the algorithms, it would be possible to comprehend the behavior of network members in real-world situations. Community detection techniques are useful for identifying individuals with similar interests and keeping them connected, allowing the extraction of groups with similar characteristics. The dataset used was collected from Twitter during the pre, during, and post-2022 World Cup period. Graphs are used in this analysis to take advantage of properties like communities and the identification of active and influential individuals through the edges connecting vertices. To achieve the goal, the initial step involves characterizing the collected data sample to understand the data and obtain a summary of its main characteristics. The ultimate outcome of this work is the identification and analysis of the communities formed within this context.

Keywords: Data analysis. Community detection. Twitter. World Cup.

LISTA DE FIGURAS

Figura 1 – Volume de tweets durante o período coletado	31
Figura 2 – Nuvens das 100 palavras mais populares dos tweets e retweets	32
Figura 3 – Popularidade das palavras-chave ao longo das semanas	33
Figura 4 – Distribuições dos números de tweets e retweets por tipo de conta	34
Figura 5 – Top-10 emojis mais frequentes em tweets e retweets	36
Figura 6 – Evolução semanal da diferença entre a fração de sentimentos positivos e negativos	38
Figura 7 – Top LIWC atributos extraídos dos tweets e retweets coletados	39
Figura 8 – Exemplo de grafo com formato estrela	42
Figura 9 – Distribuição de número de indivíduos por comunidade nos dia da cerimônia de abertura	44
Figura 10 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da cerimônia de abertura	46
Figura 11 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da cerimônia de abertura	47
Figura 12 – Análise de psicolinguística das top 5 maiores comunidades do dia da cerimônia de abertura	48
Figura 13 – Distribuição de número de indivíduos por comunidade nos dia da estreia de seleção brasileira	51
Figura 14 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da estreia da seleção brasileira	54
Figura 15 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da estreia da seleção brasileira	55
Figura 16 – Análise de psicolinguística das top 5 maiores comunidades do dia da estreia da seleção brasileira	56
Figura 17 – Distribuição de número de indivíduos por comunidade nos dia da eliminação da seleção brasileira	60
Figura 18 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da eliminação da seleção brasileira	62
Figura 19 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da eliminação da seleção brasileira	63
Figura 20 – Análise de psicolinguística das top 5 maiores comunidades do dia da eliminação da seleção brasileira	64
Figura 21 – Distribuição de número de indivíduos por comunidade nos dia da partida final	68
Figura 22 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da partida final	70

Figura 23 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da partida final	71
Figura 24 – Análise de psicolinguística das top 5 maiores comunidades do dia da partida final	72

LISTA DE TABELAS

Tabela 1 – Comparação de redes sociais, métodos de pesquisa e principais achados dos trabalhos	21
Tabela 2 – Principais estatísticas do conjunto de dados	30
Tabela 3 – Top 5 contas mais ativas em # de tweets e retweets: contas verificadas	35
Tabela 4 – Top 5 contas mais ativas em # de tweets e retweets: contas não verificadas	35
Tabela 5 – Exemplos do uso de emojis	37
Tabela 6 – Informações gerais sobre os grafos do dia da cerimônia de abertura . .	43
Tabela 7 – Número de internautas e retweets por comunidade no dia da cerimônia de abertura	44
Tabela 8 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da cerimônia de abertura	45
Tabela 9 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da cerimônia de abertura	45
Tabela 10 – Análise de tópicos LDA para a comunidade 1 do dia da cerimônia de abertura	49
Tabela 11 – Análise de tópicos LDA para a comunidade 2 do dia da cerimônia de abertura	49
Tabela 12 – Análise de tópicos LDA para a comunidade 3 do dia da cerimônia de abertura	50
Tabela 13 – Análise de tópicos LDA para a comunidade 4 do dia da cerimônia de abertura	50
Tabela 14 – Análise de tópicos LDA para a comunidade 5 do dia da cerimônia de abertura	51
Tabela 15 – Informações gerais sobre os grafos do dia da estreia da seleção brasileira	51
Tabela 16 – Número de internautas e retweets por comunidade no dia da estreia da seleção brasileira	52
Tabela 17 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da estreia da seleção brasileira	52
Tabela 18 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da estreia da seleção brasileira	52
Tabela 19 – Análise de tópicos LDA para a comunidade 1 do dia da estreia da seleção brasileira	57
Tabela 20 – Análise de tópicos LDA para a comunidade 2 do dia da estreia da seleção brasileira	57
Tabela 21 – Análise de tópicos LDA para a comunidade 3 do dia da estreia da seleção brasileira	58

Tabela 22 – Análise de tópicos LDA para a comunidade 4 do dia da estreia da seleção brasileira	58
Tabela 23 – Análise de tópicos LDA para a comunidade 5 do dia da estreia da seleção brasileira	59
Tabela 24 – Informações gerais sobre os grafos do dia da eliminação da seleção brasileira	59
Tabela 25 – Número de internautas e retweets por comunidade no dia da eliminação da seleção brasileira	60
Tabela 26 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da eliminação da seleção brasileira	61
Tabela 27 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da eliminação da seleção brasileira	61
Tabela 28 – Análise de tópicos LDA para a comunidade 1 do dia da eliminação da seleção brasileira	65
Tabela 29 – Análise de tópicos LDA para a comunidade 2 do dia da eliminação da seleção brasileira	66
Tabela 30 – Análise de tópicos LDA para a comunidade 3 do dia da eliminação da seleção brasileira	66
Tabela 31 – Análise de tópicos LDA para a comunidade 4 do dia da eliminação da seleção brasileira	66
Tabela 32 – Análise de tópicos LDA para a comunidade 5 do dia da eliminação da seleção brasileira	67
Tabela 33 – Informações gerais sobre os grafos do dia da partida final	67
Tabela 34 – Número de internautas e retweets por comunidade no dia da final do evento	68
Tabela 35 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da partida final	69
Tabela 36 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da partida final	69
Tabela 37 – Análise de tópicos LDA para a comunidade 1 do dia da partida final	73
Tabela 38 – Análise de tópicos LDA para a comunidade 2 do dia da partida final	74
Tabela 39 – Análise de tópicos LDA para a comunidade 3 do dia da partida final	75
Tabela 40 – Análise de tópicos LDA para a comunidade 4 do dia da partida final	75
Tabela 41 – Análise de tópicos LDA para a comunidade 5 do dia da partida final	76

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Analysis of Variance
API	Application Programming Interface
FIFA	Fédération Internationale de Football Association
LDA	Latent Dirichlet Allocation
LeIA	Léxico para Inferência Adaptada
LIWC	Linguistic Inquiry and Word Count
VADER	Valence Aware Dictionary and sEntiment Reasoner

SUMÁRIO

1	INTRODUÇÃO	14
1.1	OBJETIVO GERAL	15
1.2	OBJETIVOS ESPECÍFICOS	15
1.3	METODOLOGIA	16
1.4	ESTRUTURA DO TRABALHO	17
2	TRABALHOS RELACIONADOS	18
2.1	REDES DE OPINIÃO EM REDES SOCIAIS	18
2.2	USO DO TWITTER PARA DISCUSSÃO DE EVENTOS REAIS	19
2.3	REDES SOCIAIS E EDIÇÕES DA COPA DO MUNDO	20
2.4	ANÁLISE COMPARATIVA	21
3	FUNDAMENTAÇÃO TEÓRICA	22
3.1	TWITTER	22
3.2	CARACTERIZAÇÃO DE DADOS	22
3.2.1	Métodos de Caracterização de Dados	23
3.3	GRAFO	25
3.3.1	Comunidade	26
3.4	DETECÇÃO DE COMUNIDADES	26
3.4.1	Modularidade	26
3.4.2	Algoritmo de Louvain	27
4	CARACTERIZAÇÃO DO CONJUNTO COM TODOS OS DADOS	28
4.1	TIPOS DE ANÁLISES REALIZADAS	28
4.2	CONJUNTO DE DADOS	29
4.2.1	Análise Exploratória do Conjunto de Dados	30
4.3	RESULTADOS E ANÁLISES	33
4.3.1	Perfil dos Indivíduos	33
4.3.2	Análise dos Emojis	36
4.3.3	Análise de Sentimentos	37
4.3.4	Análise Psicolinguística	39
5	ANÁLISE E DETECÇÃO DE COMUNIDADES	41
5.1	TIPOS DE ANÁLISES REALIZADAS	41
5.2	DIA DA CERIMÔNIA DE ABERTURA DA COPA DO MUNDO DE 2022	43
5.3	DIA DA ESTREIA DA SELEÇÃO BRASILEIRA NA COMPETIÇÃO	51
5.4	DIA DA ELIMINAÇÃO DA SELEÇÃO BRASILEIRA NA COMPETIÇÃO	59
5.5	DIA DA PARTIDA FINAL DA COPA DO MUNDO DE 2022	67
5.6	ANÁLISE COMPARATIVA	76
6	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	78
	REFERÊNCIAS	80

APÊNDICE A – CÓDIGO	84
APÊNDICE B – ARTIGO SBC	111

1 INTRODUÇÃO

As redes sociais são plataformas utilizadas pela população para expor seus pensamentos, encontrar pessoas que compartilham das mesmas ideias e gerar discussões. Pelo grande número de internautas e por sua utilização diária, os mais diversos temas são abordados, desde pequenos eventos do dia a dia a eventos maiores e polêmicos que envolvem um grande número de pessoas. Sendo assim, se torna um importante meio para a difusão de informações e pode ser utilizado com esse propósito, substituindo canais exclusivos de notícias ou jornais, por exemplo. Apesar de ser positivo para informar a população dos acontecimentos, também pode ser prejudicial por gerar grande alcance de informações falsas e polarizar discussões. Conforme o relatório da visão geral global, atualmente, a população total mundial é de 7,8 bilhões de pessoas, e destas, 4,2 bilhões são usuárias de redes sociais. Em um ano o número de pessoas que utilizam redes sociais aumentou 490 milhões, ou seja, um crescimento de 13%. Isto significa que 53% da população mundial utilizam redes sociais (KEMP, 2021).

Ao longo dos últimos anos, a ampla utilização das redes sociais gerou um aumento na quantidade de influenciadores e impulsionou o conteúdo gerado por eles, sendo muitas destas pessoas inicialmente não famosas. Como consequência, estas pessoas passaram a atrair uma multidão de seguidores, muitas vezes formando uma rede que compartilham das mesmas opiniões ou que debatem sobre elas. Essas redes alcançam muitos indivíduos tão rapidamente que, em pouco tempo, atingem âmbito nacional e até mesmo mundial. Estudá-las contribui para uma percepção maior de como a internet e as redes sociais, como um meio rápido e de fácil acessibilidade, facilitam essa influência dos formadores de opinião sobre a população. O Twitter é uma das 10 redes sociais mais populares no Brasil e uma das 15 mais utilizadas em todo o mundo (VOLPATO, 2023), o que o torna um importante meio de difusão de informações e, conseqüentemente, uma ótima fonte de dados, permitindo a extração de opiniões de grande parte da população.

Nas redes sociais, interações tendem a ser agrupadas em comunidades (COELHO *et al.*, 2013). As comunidades são muito comuns em redes de informações, podendo existir várias. Os nodos dentro de uma comunidade costumam ser densamente conectados e podem se sobrepôr, ou seja, participar de diversas comunidades. Isso ocorre porque em uma rede social pode existir interações com amigos, colegas, familiares e com pessoas aleatórios da plataforma, formando uma comunidade muito densa. As ligações entre comunidades são chamadas de pontes, responsáveis pelo pertencimento das comunidades a uma rede, diferenciando de uma ilha de nodos.

Existem diferentes algoritmos para detecção de comunidades em redes, como louvain, surprise e leiden (GIRVAN; NEWMAN, 2002). A detecção de comunidade pode auxiliar mídias sociais a identificarem pessoas com interesses comuns e mantê-las firmemente conectadas. Também podem ser usadas no aprendizado de máquina para detectar

grupos com propriedades semelhantes e extraí-los para, por exemplo, para identificar manipuladores dentro de uma rede social.

Este trabalho tem como objetivo analisar este fenômeno de disseminação de informações nas redes sociais, onde grandes redes de opiniões são formadas e as postagens de alto alcance atingem pessoas dos mais variados lugares, idades e classes sociais. O conjunto de dados utilizados neste trabalho foi disponibilizado pelo projeto "PROCORES: Caracterização e Modelagem de Processos de Contágio em Redes Sociais de Diferentes Domínios"¹ (MALAGOLI *et al.*, 2021). Os dados foram extraídos do Twitter, no qual o tema da coleta foi a Copa do Mundo de futebol da FIFA, que ocorreu em 2022 no Catar. Este é um evento de grande importância e paixão para grande parte dos brasileiros, pois desperta um intenso fervor patriótico e une o país em torno do esporte, uma vez que o futebol é profundamente enraizado na cultura brasileira, sendo considerado um símbolo de identidade nacional e orgulho (PENFOLD, 2019). A análise desses dados permite compreender as discussões, opiniões e diferentes perspectivas dos brasileiros sobre esse evento esportivo de grande impacto no país. Foram analisados quase 12 milhões de tweets coletados entre novembro e dezembro de 2022, cobrindo o período pré, durante e pós evento. A caracterização destes dados examina a divulgação de informações considerando dois aspectos cruciais: o envolvimento dos internautas e as propriedades dos conteúdos.

Além de explorar as conversas relacionadas ao futebol, este estudo também busca compreender como os brasileiros utilizam o Twitter para expressar opiniões, compartilhar informações e participar de discussões sobre temas relevantes relacionados à Copa do Mundo de 2022. Por meio do conceito de computação social, que estuda a interação entre pessoas e tecnologia, foi possível analisar a participação dos indivíduos nas redes sociais e a sua influência na disseminação de opiniões e informações durante o evento. Através dessa abordagem, foram analisadas e detectadas redes de opiniões por semelhança e aplicados algoritmos de detecção de comunidade.

1.1 OBJETIVO GERAL

Este Trabalho de Conclusão de Curso tem como objetivo geral a caracterização e análise de formação de comunidades identificadas através de um algoritmo de detecção de comunidades no contexto da discussão dos brasileiros sobre a Copa do Mundo de 2022 no Twitter.

1.2 OBJETIVOS ESPECÍFICOS

Tendo em vista o objetivo geral descrito acima, são identificados os seguintes objetivos específicos:

¹ <https://procores.com.br/>

- Caracterização do conjunto de dados;
- Análise exploratória do conjunto de dados, utilizando métodos como análise de nuvens de palavras, análise da popularidade das palavras chave, análise de sentimentos, análise psicolinguísticas etc;
- Modelagem, através de grafos, das postagens comuns entre os internautas. No contexto do Twitter, estas postagens comuns podem ser definidas através da repostagem de um conteúdo exatamente igual, por exemplo;
- Uso de algoritmo de detecção de comunidades considerando os grafos modelados;
- Caracterização e comparação do conteúdo das comunidades detectadas.

1.3 METODOLOGIA

A metodologia adotada para a obtenção dos resultados é composta por quatro passos. O passo de implementação é a análise ou verificação dos resultados alcançados e dados coletados, onde são utilizados os indicadores de acompanhamento, através de métricas de avaliação de resultados, descritos no Capítulo 5. Este passo pode ocorrer simultaneamente à realização do projeto, momento em que se verifica se o trabalho está sendo feito da forma devida, ou após a execução quando são feitas análises estatísticas dos dados e verificação dos itens de controle. Nesta fase podem ser detectados erros ou falhas.

1. **Levantamento bibliográfico:** o levantamento bibliográfico trata da busca por trabalhos sobre o assunto investigado, como forma de verificar os trabalhos desenvolvidos recentemente, ou em desenvolvimento, por outros grupos de pesquisa. Esta etapa proporciona o amadurecimento dos temas pesquisados e seu posterior estudo analítico. O Capítulo 2 descreve alguns trabalhos.
2. **Proposta da solução:** com base no levantamento bibliográfico realizado, levando-se em conta oportunidades encontradas, é especificada a proposta de solução, que é composta pelas seguintes tarefas:
 - Caracterização do conjunto com todos os dados: processo de análise dos dados que envolve a exploração e descrição das características, padrões e propriedades dos dados disponíveis no contexto da Copa do Mundo de 2022;
 - Detecção de comunidades: estruturação dos dados no formato de grafo e execução de um algoritmo de detecção de comunidades;
 - Caracterização das comunidades: mesmo processo da caracterização do conjunto com todos os dados, porém limitando aos dados de cada comunidade.

3. **Implementação da solução:** todas as tarefas planejadas na proposta são implementadas para possibilitar, posteriormente, a análise da solução através da validação de resultados.
4. **Validação e ajuste das propostas:** ao final da tarefa de implementação, são realizadas análises para validação da proposta e, se necessário, a realização de ajustes. O objetivo principal é confirmar se os requisitos específicos para um determinado objetivo foram cumpridos.

1.4 ESTRUTURA DO TRABALHO

Este Capítulo apresenta a introdução, o objetivo geral e os objetivos específicos. O Capítulo 2 traz trabalhos relacionados a este projeto. No Capítulo 3 os conceitos fundamentais para a compreensão deste trabalho são apresentados. No Capítulo 4 encontra-se a caracterização do conjunto de dados completo. O Capítulo 5 apresenta análises das comunidades detectadas. Por fim, o Capítulo 6 apresenta as considerações finais e os trabalhos futuros.

2 TRABALHOS RELACIONADOS

Neste trabalho é abordada a identificação de comunidades em redes sociais, mais especificamente no Twitter. O objetivo principal é analisar como informações sobre a Copa do Mundo de 2022 foram difundidas entre os internautas a partir da análise de dados coletados dessa rede. Este capítulo apresenta alguns artigos relacionados às análises realizadas neste trabalho. Os artigos são divididos em três grupos principais: Redes de Opinião nas Redes Sociais, Análise de Dados do Twitter e Copas do Mundo.

2.1 REDES DE OPINIÃO EM REDES SOCIAIS

Em Belegante e Menezes (2015), os autores investigaram a influência dos formadores de opinião nas redes sociais (como Orkut, Facebook, MySpace, Twitter e LinkedIn) e examinaram como o público reage a isso. Um grande número de indivíduos alcançou notoriedade ao expressar suas opiniões em redes sociais, tornando-se formadores de opinião que exercem impacto sobre uma audiência diversificada em termos de localização, faixa etária e classe social. Para que o objetivo do estudo de identificar o processo de desenvolvimento dos formadores fosse alcançado, foi utilizado o método de pesquisa bibliográfica, recorrendo a autores secundários que abordam a interatividade e a influência nas redes sociais, a fim de fundamentar as análises apresentadas. A pesquisa explora temas relacionados à influência e à manipulação nas redes sociais, analisando o papel dos formadores de opinião e das empresas na formação das perspectivas das pessoas, entre outros tópicos relevantes para uma explicação clara e objetiva.

O estudo de Teixeira e Azevedo (2011) empregou técnicas de Análise de Sentimentos para avaliar se as informações presentes em duas plataformas de redes sociais (Facebook e Twitter) poderiam ser empregadas para estimar valores associados à comercialização de produtos ou serviços a serem lançados no mercado. O uso cotidiano crescente das redes sociais trouxe consigo novas funcionalidades e aplicações significativas. Este aumento recente de popularidade nesse tipo de serviço tem proporcionado diversas oportunidades. Os internautas contribuem ativamente com suas opiniões e conhecimentos, formando assim um vasto repositório de informações. Empresas têm reconhecido cada vez mais o valor dessas redes sociais, enxergando nelas uma maneira eficaz de promover seus produtos junto ao público e analisar como esses produtos são percebidos.

O trabalho de Abbade, Della Flora e Noro (2014) teve como propósito analisar a postura de estudantes universitários em relação à influência interpessoal nas redes sociais virtuais durante o processo de tomada de decisão de consumo. Para atingir esse objetivo, foi realizado um levantamento (survey) envolvendo 200 estudantes de uma Instituição de Ensino Superior (IES) localizada em Santa Maria, Rio Grande do Sul. A seleção da amostra foi realizada por adesão, e a coleta de dados ocorreu em um ambiente virtual. Foram ajustadas escalas para mensurar e avaliar a disposição dos estudantes universitários

para influenciar e serem influenciados por seus contatos em redes sociais virtuais. Os resultados indicam que as escalas adaptadas são eficazes para medir os aspectos propostos. Além disso, observou-se que os homens tendem a ter uma maior capacidade de influenciar as opiniões de seus contatos em redes sociais virtuais. O tempo de acesso à internet também influencia de maneira positiva e significativa a propensão dos indivíduos para serem influenciados por seus contatos nessas redes. A correlação entre a capacidade de influenciar e a propensão para ser influenciado é significativa e positiva.

2.2 USO DO TWITTER PARA DISCUSSÃO DE EVENTOS REAIS

No trabalho de Malagoli *et al.* (2021) foi investigada a percepção do público sobre a vacinação contra a COVID-19 no Twitter. Foram analisados mais de 9 milhões de tweets em português, em um período de dois meses correspondentes aos estágios iniciais da vacinação no Brasil e no mundo. Os resultados fornecem um entendimento inicial sobre a dinâmica do debate online sobre a vacinação contra a COVID-19, evidenciando como as pessoas usam o mundo online para compartilhar suas impressões e preocupações sobre o assunto.

No trabalho de Paiva *et al.* (2023), foi investigada a percepção do público em relação a temas relacionados ao feminismo no Twitter. Foram analisando mais de 700 mil tweets em português nos períodos antes, durante e após as eleições. Os resultados oferecem um entendimento inicial sobre a dinâmica do debate online em torno de temas sensíveis, destacando como as pessoas utilizam o ambiente online para compartilhar suas opiniões, impressões e preocupações em relação a assuntos relacionados ao feminismo durante o período analisado.

Em Araujo *et al.* (2023) foi investigada a promoção coordenada de campanhas de propaganda política antecipadas realizadas por pessoas do Twitter, com foco no período pré-eleitoral brasileiro de 2022. A metodologia explorada envolve a modelagem de uma rede baseada em co-retweets, a extração de um backbone da rede e, por fim, a identificação e análise de comunidades, com foco em características dos indivíduos e do conteúdo compartilhado. Os resultados revelam um número significativo de comunidades que promovem conteúdo relacionado a diversos pré-candidatos de diferentes espectros políticos, incluindo direita e esquerda. Além disso, foi constatado que as comunidades de direita são muito maiores em comparação com as de esquerda, tanto em relação ao número de indivíduos quanto ao volume de informações compartilhadas. Os resultados podem fornecer insights interessantes para o entendimento do fenômeno no contexto brasileiro, bem como, futuramente, para auxiliar a formulação de mecanismos que sejam eficientes na detecção de campanhas eleitorais antecipadas em plataformas sociais.

2.3 REDES SOCIAIS E EDIÇÕES DA COPA DO MUNDO

O trabalho de Lins (2020) tem como objetivo verificar quais as variáveis que influenciam o comportamento de compra por impulso de acessórios de torcida durante megaeventos esportivos. Estes eventos são momentos em que o espírito de nacionalismo costuma ser elevado, fazendo com que as pessoas adquiram produtos para demonstrar apoio ao seu país na competição. Participaram do estudo 441 brasileiros, sendo 264 mulheres e 177 homens, com média de idade de 34.56 anos e desvio padrão de 15.89 anos. A coleta de dados foi realizada através de um questionário online, que foi divulgado através do Facebook, WhatsApp e listas de e-mails durante a Copa do Mundo de Futebol FIFA 2018. Os resultados mostraram que os homens são mais fanáticos pela copa e tendem a se envolver mais no evento. Um modelo de mediação sequencial, aplicado a ambos os sexos, revelou que quanto mais forte é a identidade nacional, maior é o fanatismo, que leva a um maior envolvimento, resultando na compra por impulso de acessórios de torcida.

O estudo de Gastaldo (2009) pesquisa a abordagem midiática em relação à Copa do Mundo no Brasil. Em uma análise crítica sobre a formação social dos "meios de comunicação de massa", foi examinado como a mídia influencia a sociedade, explorando o conceito de "mediação". Ao utilizar dados sobre a audiência e os temas predominantes na televisão brasileira durante a Copa do Mundo de 1998, foi investigado o papel desempenhado pela mídia na construção do interesse social pela competição no Brasil. Embora não tenha sido buscado negar a legitimidade do interesse coletivo pelo futebol no país, a contribuição significativa da mídia nesse processo se destaca.

O trabalho de Filho e Silva (2014) tem como objetivo fazer a coleta, estruturação e descoberta de conhecimento de dados textuais extraídos do Twitter, para mapear a opinião dos internautas sobre Copa do Mundo de 2014. Os sentimentos das postagens, popularmente conhecido como tweets, são categorizadas neste trabalho como: positivo, negativo, ambíguo ou neutro. Os resultados apresentados no artigo mostram que o aumento das redes sociais nos últimos anos permitiu as pessoas se conectarem e compartilharem informações em tempo real com milhares de outras pessoas em um curto espaço de tempo. É comum a postagem em redes sociais de opiniões a respeito de acontecimentos como: grandes eventos, lançamentos de produtos, catástrofes e epidemias. Assim, foi possível perceber que acompanhar o que está sendo discutido nas redes sociais pode ser um diferencial para as organizações que desejam elaborar melhores estratégias de marketing e obter feedback sobre algum produto ou evento. No entanto, essa quantidade de dados continua crescendo e a análise desses dados de forma não automatizada pode ser um problema não trivial.

2.4 ANÁLISE COMPARATIVA

A partir dos trabalhos apresentados acima, foi possível construir a Tabela 1 com as redes sociais abordadas, métodos de pesquisa e principais achados. Observa-se que grande parte dos trabalhos utilizaram o Twitter como rede social consultada, porém outras redes sociais foram utilizadas. Quanto aos métodos de pesquisa, muitos utilizam formas que trabalham com análise de um conjunto de dados, já outros fazem o uso de questionários com o público alvo ou trabalham apenas com pesquisa bibliográfica. Na coluna de principais achados estão descritos os principais aspectos que os trabalhos agregaram para este presente trabalho.

Tabela 1 – Comparação de redes sociais, métodos de pesquisa e principais achados dos trabalhos

Autor(es)	Ano	Título do trabalho	Rede social abordada	Método de pesquisa	Principais achados
Belegante e Menezes	2015	A influência dos formadores de opinião nas redes sociais	Orkut, Facebook, MySpace, Twitter e LinkedIn	Pesquisa bibliográfica	Contextualização sobre os formadores de opinião nas redes sociais
Teixeira e Azevedo	2011	Análise de opiniões expressas nas redes sociais	Facebook e Twitter	Análise de sentimentos com dados extraídos do Facebook e Twitter.	Técnicas de análise em dados extraídos de redes sociais
Abbate, Della Flora e Noro	2014	A Influência Interpessoal em Redes Sociais Virtuais e as Decisões de Consumo	Blogs, Facebook, Fóruns de Discussões, Orkut, Twitter, LinkedIn, etc.	Questionário online	Influência das redes sociais em eventos externos
Malagoli, Stancioli, Ferreira, Vasconcelos, Silva e Almeida	2021	Caracterização do debate no Twitter sobre a vacinação contra a COVID-19 no Brasil	Twitter	Extração e análise de dados do Twitter	Caracterização e análise de dados extraídos do Twitter
Paiva, Barbosa, Silva e Moro	2023	O debate do feminismo no Twitter: Um estudo de caso das eleições brasileiras de 2022	Twitter	Extração e análise de dados do Twitter	Caracterização e análise de dados extraídos do Twitter
Araujo, Ferreira, Reis, Silva e Almeida	2023	Identificação e Caracterização de Campanhas de Propagandas Eleitorais Antecipadas Brasileiras no Twitter	Twitter	Modelagem e backbone de rede e identificação e análise de comunidades	Caracterização e análise de comunidades identificadas no Twitter
Lins	2020	Preparando-se para a Copa do Mundo: o que leva os brasileiros a comprar impulsivamente produtos para apoiar o seu país?	Não se aplica	Questionário online	Contextualização do fanatismo brasileiro pela Copa do Mundo
Gastaldo	2009	"O país do futebol" mediatizado: mídia e Copa do Mundo no Brasil	Não se aplica	Investigação dos dados de audiência e temas predominantes na televisão brasileira	Influência da mídia no contexto da Copa do Mundo no Brasil
Filho e Silva	2014	Mineração de textos: análise de sentimentos utilizando Tweets referentes à Copa do Mundo 2014	Twitter	Extração, estruturação e análise de sentimentos de dados do Twitter.	Técnicas de análise de dados extraídos do Twitter sobre a Copa do Mundo
Presente trabalho	2023	Caracterização e análise de formação de comunidades no contexto da Copa do Mundo de 2022	Twitter	Análise de dados do Twitter	Não se aplica

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, foram tratados alguns conceitos fundamentais para a compreensão deste trabalho.

3.1 TWITTER

O Twitter (atualmente X) ¹ é uma plataforma de mídia social amplamente utilizada no Brasil, sendo uma das 10 redes sociais mais populares no país e uma das 15 mais utilizadas em todo o mundo (VOLPATO, 2023). Foi lançado em 2006 e possui atualmente uma base de quase 556 milhões de contas no mundo.

O Twitter é uma rede social baseada em microblogs, onde as pessoas podem publicar mensagens com até 280 caracteres. Essas mensagens são chamadas de tweets e oferecem suporte para a inclusão de mídia, como vídeos, imagens e links, além do uso de hashtags, que são palavras-chave precedidas pelo símbolo # e amplamente utilizadas para filtrar ou promover conteúdos relevantes. Além disso, as pessoas têm a possibilidade de curtir e compartilhar tweets de outras pessoas, sendo essa última funcionalidade conhecida como retweet.

Em novembro de 2022, o empresário Elon Musk comprou o Twitter. Após a compra foram feitas algumas mudanças na rede social, como o nome para X, mas essas alterações não impactaram na coleta dos dados utilizados neste trabalho por terem começado apenas no ano de 2023 ².

3.2 CARACTERIZAÇÃO DE DADOS

A caracterização de dados é um processo na análise de dados que envolve a exploração e descrição das características, padrões e propriedades dos dados disponíveis em um determinado contexto. Ela permite uma compreensão mais profunda do problema em questão, ajudando a identificar insights relevantes e responder a perguntas específicas relacionadas aos dados. Seus objetivos são amplos e variam dependendo das necessidades e do contexto do projeto em questão.

Inicialmente, a caracterização busca compreender a natureza dos dados, examinando sua estrutura, qualidade e propriedades. Essa compreensão é fundamental para realizar análises mais aprofundadas.

Além disso, a caracterização de dados tem como objetivo identificar padrões e tendências significativas nos dados. Por meio da análise de estatísticas descritivas é possível compreender a distribuição dos dados e identificar informações relevantes. Técnicas de visualização de dados podem ser aplicadas, como gráficos e histogramas, para identificar

¹ <https://x.com/>

² <https://gauchazh.clicrbs.com.br/tecnologia/noticia/2023/07/de-twitter-a-x-as-mudancas-na-rede.html>

tendências ao longo do tempo, sazonalidades ou anomalias nos dados. Outro objetivo da caracterização de dados é trazer ideias de técnicas de análise apropriadas, pois ao compreender as características dos dados é possível escolher e adaptar os métodos mais adequados para o problema em questão.

Por fim, ao compreender as características dos dados, os padrões encontrados e as limitações associadas, é possível fornecer informações valiosas para apoiar nas tomadas de decisões.

3.2.1 Métodos de Caracterização de Dados

Nessa seção estão descritos alguns conceitos relacionados aos métodos de análise de dados utilizados para a etapa de caracterização, conforme abaixo:

- **Nuvem de Palavras:**

Nuvem de palavras é uma ferramenta visual que mostra a frequência das palavras usadas em um texto. Através de algoritmos, é possível identificá-las e criar imagens com dezenas de palavras, onde o tamanho indica a frequência ou relevância dessas em meio a centenas ou milhares de postagens (VASCONCELLOS-SILVA; ARAÚJO-JORGE, 2019). A criação das imagens no formato de nuvem de palavras deste trabalho foram geradas pela ferramenta *WordCloud*³ do python.

- **LeIA:**

O Léxico para Inferência Adaptada (LeIA) é uma versão adaptada do léxico e ferramenta de análise de sentimentos chamada Valence Aware Dictionary and sEntiment Reasoner (VADER)⁴, especificamente desenvolvida para lidar com textos em português. Seu objetivo principal é analisar os sentimentos expressos em textos provenientes de mídias sociais, embora também possa ser aplicado a textos de outros domínios. Uma das principais vantagens do LeIA é que a mesma API do VADER é mantida, o que significa que não é necessário realizar um pré-processamento especial no texto de entrada antes de utilizá-lo. Assim, o processo de análise de sentimentos é simplificado, permitindo que obtenha-se resultados de maneira rápida e direta. Após realizar a análise de sentimentos, o LeIA fornece como saída um dicionário contendo quatro campos principais: *pos*, *neg*, *neu* e *compound*. O campo *pos* representa a porcentagem de sentimento positivo no texto, o campo *neg* a porcentagem de sentimento negativo e o campo *neu* a porcentagem de sentimento neutro. Por fim, o campo *compound* é um valor numérico que normaliza o sentimento geral do texto, variando de -1 (extremamente negativo) a +1 (extremamente positivo). O valor *compound* é particularmente útil para descrever o sentimento predominante no

³ <https://pypi.org/project/wordcloud/>

⁴ <https://pypi.org/project/vaderSentiment/>

texto analisado. Valores maiores ou iguais a 0.05 indicam um sentimento positivo, valores menores ou iguais a -0.05 representam um sentimento negativo, enquanto valores entre -0.05 e 0.05 sugerem um sentimento neutro. Em resumo, o LeIA é uma ferramenta adaptada do VADER para a língua portuguesa, permitindo a análise de sentimentos em textos expressos em mídias sociais, assim como em outros domínios. Sua facilidade de uso e a capacidade de preservar a API do VADER tornam-no uma opção conveniente para aqueles que desejam compreender os sentimentos presentes em textos em português.

- **LIWC:**

O Linguistic Inquiry and Word Count (LIWC) é um software de análise de texto que permite calcular o grau de utilização de diferentes categorias de palavras em uma ampla variedade de textos. Essa ferramenta baseia-se em um dicionário léxico conhecido como dicionário LIWC. Recentemente, uma versão específica para a língua portuguesa foi disponibilizada, chamada de *Brazilian Portuguese LIWC 2007 Dictionary*⁵. Com o LIWC, é possível realizar análises detalhadas sobre o uso de palavras relacionadas a diferentes categorias, como emoções, temas, estilo de escrita e até mesmo a análise de sentimentos (BALAGE FILHO; PARDO; ALUISIO, 2013).

- **Teste de Kruskal-Wallis:**

O teste de Kruskal-Wallis é uma extensão do teste de Wilcoxon-Mann-Whitney, usado para comparar a distribuição de duas amostras. No entanto, o teste de Kruskal-Wallis permite a comparação de três ou mais grupos em amostras independentes, tornando-se uma alternativa não paramétrica à Analysis of Variance (ANOVA) para um fator. A principal diferença entre o teste de Kruskal-Wallis e a ANOVA reside nas suposições necessárias para cada um. Enquanto a ANOVA requer validação das suposições de normalidade, variância constante, independência dos resíduos e variáveis contínuas, o teste de Kruskal-Wallis considera apenas a independência das observações e permite variáveis contínuas ou ordinais. Além disso, enquanto a ANOVA testa a média e a variação entre os grupos, o teste de Kruskal-Wallis avalia uma pseudo-mediana. Isso significa que as diferenças nas médias de ordens ou postos são analisadas, podendo não ser necessariamente iguais às medianas dos grupos. Para realizar o teste, as observações são ordenadas em ordem crescente, independentemente dos grupos, e cada observação recebe um posto ou ordem. O menor valor recebe o posto 1, o segundo recebe o posto 2 e assim por diante, até que todas as observações sejam consideradas. Quando há observações repetidas, é recomendado atribuir o valor médio dos postos a essas observações. Isso pode resultar em um teste que indica diferença significativa entre os grupos, mesmo que as medianas sejam iguais ou próximas. Nesses casos, o teste de Kruskal-Wallis

⁵ <http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>

avalia simultaneamente a mediana e as formas de distribuição. Quando os grupos apresentam a mesma forma de distribuição de probabilidade, a interpretação do resultado pode ser baseada na mediana.

- **Coefficiente de Gini:**

O coeficiente de Gini é um índice estatístico amplamente utilizado como um indicador socioeconômico que mede a distribuição de renda em uma determinada área geográfica, como país, estado, região ou município. Esse índice permite avaliar o grau de desigualdade ou igualdade econômica em um território, utilizando uma escala que varia de 0 a 1. O valor 0 indica igualdade máxima, indicando que a renda está distribuída de forma equitativa nesse território. Quanto mais próximo de 0, maior é a igualdade na distribuição de renda. Por outro lado, um valor 1 indica desigualdade máxima, evidenciando uma concentração de renda significativa. Quanto mais próximo de 1, maior é a desigualdade no território em questão. Neste estudo, o coeficiente de Gini é utilizado para realizar uma análise psicolinguística, sendo essencial para identificar os atributos mais discrepantes da análise LIWC realizada.

- **Algoritmo LDA**

O algoritmo Latent Dirichlet Allocation (LDA) é uma técnica valiosa no processamento de linguagem natural. Ele ajuda a identificar tópicos subjacentes em coleções de documentos. O LDA é um modelo estatístico que parte do princípio de que documentos são compostos por uma mistura de tópicos não observados (BLEI; NG; JORDAN, 2001). O funcionamento do LDA é dividido em etapas. Primeiro, é preciso definir o número de tópicos desejados. Em seguida, o algoritmo atribui aleatoriamente tópicos para cada palavra em todos os documentos. Em várias iterações, o algoritmo estima a distribuição dos tópicos em cada documento e a distribuição das palavras em cada tópico. Essas estimativas são continuamente ajustadas para maximizar a probabilidade de observar os documentos. Ao final do processo, o LDA fornece a distribuição de palavras para cada tópico e a distribuição de tópicos para cada documento. O LDA é utilizado em diversas áreas, incluindo análise de sentimentos, recomendação de conteúdo e categorização de documentos. Ele oferece uma maneira poderosa de entender e extrair informações significativas de grandes volumes de textos.

3.3 GRAFO

Um grafo é uma representação abstrata de um conjunto de objetos e das relações existentes entre eles. São definidos por um conjunto de vértices e por um conjunto de arestas que ligam pares de vértices. Muitas relações entre elementos podem ser modeladas e analisadas através dessa representação.

Existem diferentes tipos de grafos, incluindo grafos direcionados, onde as arestas têm uma direção específica, e grafos não direcionados, onde as arestas não têm direção. Além disso, podem ter propriedades adicionais, como pesos atribuídos às arestas, que representam um valor associado a cada conexão (SANTIAGO, 2023).

Os grafos são representados visualmente por meio de diagramas, onde os vértices são representados por pontos e as arestas são representadas por linhas ou setas. Também podem ser representados por meio de estruturas de dados em programação, utilizando listas de adjacência, matrizes de adjacência ou outras representações mais eficientes dependendo do contexto e das necessidades específicas.

3.3.1 Comunidade

No contexto de um grafo, uma comunidade refere-se a um subconjunto de vértices do grafo que estão fortemente interconectados entre si, enquanto apresentam conexões mais fracas com os vértices fora da comunidade. Em outras palavras, uma comunidade é um grupo de vértices que exibe alta densidade de conexões internas e baixa densidade de conexões externas. Além disso, um vértice pode fazer parte de diversas comunidades, estabelecendo assim uma ligação entre elas (GIRVAN; NEWMAN, 2002).

Em redes sociais existem comunidades extraídas a partir de grafos. Elas consistem em grupos de pessoas interconectadas, que possuem uma alta densidade de conexões internas e conexões mais fracas com pessoas fora da comunidade. Essas comunidades são formadas por pessoas com interesses, objetivos ou identidades compartilhadas, reunindo-as em torno de tópicos específicos. Um indivíduo pode participar de múltiplas comunidades, atuando como uma ponte entre elas e promovendo a interação entre diferentes grupos.

3.4 DETECÇÃO DE COMUNIDADES

A detecção de comunidades desempenha um papel crucial na análise de redes complexas, permitindo identificar agrupamentos de nós que compartilham conexões mais fortes entre si do que com o restante da rede (GIRVAN; NEWMAN, 2002). Em redes sociais, a detecção de comunidades pode ser aplicada para identificar grupos de amigos, interesses compartilhados ou até mesmo influenciadores dentro de uma plataforma.

3.4.1 Modularidade

A modularidade, frequentemente denotada por Q , é uma medida que quantifica o quão bem uma determinada divisão da rede se destaca em relação a uma configuração aleatória. Em essência, a modularidade compara a densidade de conexões dentro das comunidades com a densidade esperada de conexões em uma rede aleatória. Quanto maior a modularidade, melhor a divisão da rede em comunidades distintas.

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

Na equação 1, a variável i representa uma comunidade dentro da rede. O termo e_{ii} corresponde à fração de arestas que estão completamente contidas dentro da comunidade i , enquanto a_i representa a fração de arestas que têm pelo menos um de seus extremos pertencentes à comunidade i . A métrica de modularidade, que é denotada por um valor variando entre -1 e 1, é essencial na avaliação da qualidade das estruturas de comunidade identificadas. Quanto mais próxima de 1 for a modularidade, melhor a partição da rede em comunidades distintas.

No entanto, é importante ressaltar que o problema de determinar a partição da rede que maximiza a modularidade é conhecido como um problema NP-Completo. Isso significa que pertence a uma classe de problemas que, até o momento, não possuem solução conhecida em tempo polinomial (AIRES; NAKAMURA, 2017).

3.4.2 Algoritmo de Louvain

O Algoritmo de Louvain é uma técnica amplamente utilizada para a detecção de comunidades em redes complexas e sua aplicação está enraizada no conceito de modularidade. A implementação deste algoritmo utilizada neste trabalho é a da biblioteca NetworkX em python ⁶. Ele é um algoritmo heurístico que busca encontrar a partição da rede que maximize a modularidade. É realizado em duas etapas iterativas:

Na primeira etapa, conhecida como otimização local, o algoritmo aloca inicialmente cada nó em sua própria comunidade e, em seguida, itera sobre os nós da rede, movendo cada nó para a comunidade vizinha que maximiza o aumento na modularidade. Esse processo é repetido até que não seja possível melhorar a modularidade. Na segunda etapa, chamada otimização global, as comunidades identificadas na fase anterior são tratadas como nós em um novo grafo, onde as arestas entre os novos vértices representam a soma do peso das arestas entre as comunidades. O algoritmo repete o processo de otimização local nesse novo grafo, novamente visando maximizar a modularidade.

Essas duas etapas se alternam até que não seja mais possível melhorar a modularidade. O resultado final é uma partição da rede em comunidades que, teoricamente, representa a estrutura subjacente da rede de maneira mais eficaz (AIRES; NAKAMURA, 2017).

⁶ <https://pypi.org/project/networkx/>

4 CARACTERIZAÇÃO DO CONJUNTO COM TODOS OS DADOS

Neste capítulo foram apresentadas as principais análises considerando os quase 12 milhões de tweets/retweets coletados. Na seção 4.2 uma visão geral do conjunto de dados foi apresentada, enquanto na seção 4.3 os resultados das análises conduzidas foram detalhados, abrangendo análises de perfil dos indivíduos, emojis, sentimentos e aspectos da psicolinguística.

4.1 TIPOS DE ANÁLISES REALIZADAS

A caracterização dos dados é uma etapa baseada em análises realizadas para maior compreensão dos dados. Inicialmente, através de uma análise geral, foi verificado o volume de dados coletados ao longo do tempo, examinando nuvens de palavras e avaliando a popularidade das palavras-chave utilizadas na coleta. Além disso, foram realizadas análises de perfil dos indivíduos, de emojis, de sentimentos e psicolinguística.

A análise de perfil dos indivíduos foi examinada em relação a distribuição de tweets e retweets de acordo com o tipo de conta. Foram analisados os 10 emojis mais frequentes em tweets e retweets, fornecendo exemplos de como aparecem. Também foram feitas análises de sentimentos e psicolinguística.

Na análise de sentimento foi utilizada a ferramenta *LeIA*¹, que extrai o sentimento de cada tweet. O *LeIA* fornece uma pontuação inteira que descreve o sentimento predominante no texto. Foram considerados tweets e retweets com pontuações menores que -0,05 como negativos, maiores que 0,05 como positivos e entre -0,05 e 0,05 como neutros. O *LeIA* é uma ferramenta muito utilizada para análise de sentimentos em textos de redes sociais, mas ele também pode ser aplicado a textos de outros domínios. A fim de entender como o sentimento expresso pelas pessoas varia, foi utilizado um mapa de calor para mostrar o contraste dos sentimentos expressos nos tweets e retweets através da pontuação contrastiva, sendo calculada como a diferença entre a fração de tweets positivos e negativos.

Na análise psicolinguística foi utilizado o léxico LIWC² (TAUSCZIK; PENNEBAKER, 2010) para categorizar as palavras em diferentes atributos relacionados ao estilo linguístico, conceitos afetivos e cognitivos. A frequência média desses atributos foi calculada para cada palavra-chave nos tweets e retweets.

Nos tweets e retweets analisados, foram identificados a presença dos 64 atributos disponíveis na versão em português do LIWC de 2007. Em seguida, foram analisadas diferenças estatísticas entre os debates em torno de diferentes palavras-chave, explorando a frequência média dos atributos nos tweets e retweets associados a cada palavra-chave. O teste não paramétrico de Kruskal-Wallis (KRUSKAL; WALLIS, 1952) foi utilizado para

¹ <https://github.com/rafjaa/LeIA>

² <http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>

selecionar atributos que apresentassem diferenças significativas entre as palavras-chave. No entanto, foi observado que todos os 64 atributos apresentaram diferenças significativas.

Para lidar com a grande quantidade de atributos, o coeficiente de Gini (YITZHAKI, 1979) foi utilizado para selecionar, dentre os 64, os 20 que eram mais discriminantes. Dessa forma, foi possível identificar os atributos mais relevantes para cada palavra-chave através de um mapa de calor com esses atributos discriminantes, considerando todos os tweets e retweets. Cada célula do mapa de calor em uma coluna representa o desvio relativo de um determinado atributo para uma palavra-chave específica em relação às outras palavras-chave. As células foram coloridas em gradiente entre vermelho e azul, indicando se o atributo está acima ou abaixo da média, respectivamente. Para isso, cada coluna foi normalizada utilizando a métrica z-score, ou seja, subtraímos a média da coluna e dividimos pelo desvio padrão da coluna.

4.2 CONJUNTO DE DADOS

O conjunto de dados utilizado foi obtido dentro do contexto do projeto "PRO-CORES: Caracterização e Modelagem de Processos de Contágio em Redes Sociais de Diferentes Domínios"³. Foram coletados tweets que mencionam, ao menos uma vez, uma das palavras dentro deste conjunto de palavras-chave: *Argentina*, *BrasilNaCopa*, *Catar*, *CopaDoMundo2022*, *CopaDoMundoFIFA*, *CopaMundialFIFA*, *CopadoMundo*, *FIFA*, *FIFAWorldCup*, *Hexa*, *Messi*, *Neymar*, *Qatar2022*, *QatarWorldCup2022*, *RUMOAOHEXA*, *SelecaoBrasileira*, *Tite*, *neyday*. No total, foram coletados quase 12 milhões de tweets durante um período de 9 semanas. Essas semanas abrangeram o intervalo entre 01 de novembro e 31 de dezembro de 2022.

A Tabela 2 apresenta uma descrição inicial do conjunto de dados, mostrando os totais de tweets, retweets e internautas únicos por semana de coleta. Nas análises realizadas, considera-se tanto os tweets quanto os *replies*, que são respostas a tweets postados por outras pessoas. De maneira geral, observa-se uma tendência de crescimento no engajamento dos internautas, estimado pela frequência de postagens (tweets e retweets), na discussão dos tópicos relacionados à Copa do Mundo durante a atuação da Seleção Brasileira, sendo este da primeira à sexta semana. É possível observar que houve uma queda a partir da semana seguinte. Na oitava semana houve o crescimento de postagens e do engajamento dos internautas, uma vez que foi a semana da final do torneio.

³ <https://procores.com.br/>

Tabela 2 – Principais estatísticas do conjunto de dados

Semana	Início	#Tweets	#Retweets	#Internautas Únicos
1	01-11-2022	45393	73125	89352
2	05-11-2022	258904	318535	317448
3	12-11-2022	213547	295782	285148
4	19-11-2022	1053277	1826730	998463
5	26-11-2022	1005601	1471705	883883
6	03-12-2022	902572	1346987	912662
7	10-12-2022	556430	740736	620022
8	17-12-2022	711005	1035255	717269
9	24-12-2022	221277	253299	269587

4.2.1 Análise Exploratória do Conjunto de Dados

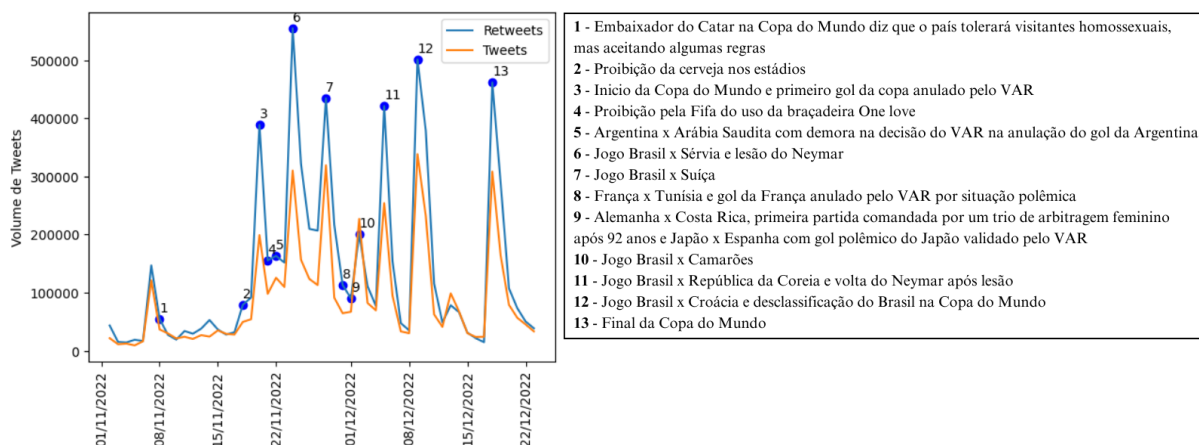
A Figura 1 apresenta uma análise mais detalhada da evolução temporal das discussões relacionadas à Copa do Mundo de 2022. Através da série temporal dos números diários de tweets e retweets, é possível observar a ocorrência de picos significativos que coincidem com eventos relevantes durante o período de coleta dos dados. Alguns desses eventos incluem os jogos da Seleção Brasileira e a proibição da cerveja nos estádios⁴.

É interessante destacar que o primeiro pico significativo ocorreu no primeiro dia da Copa do Mundo, durante a cerimônia de abertura e o início dos jogos. Os picos numerados como 6, 7, 10, 11 e 12 correspondem aos dias em que o Brasil jogou no torneio. O pico 6 é o maior deles, pois marcou a estreia da seleção brasileira e coincidiu com a lesão de Neymar⁵, um dos principais jogadores brasileiros. O pico 10 ocorreu em um jogo em que o Brasil já estava classificado para as oitavas de final, o que gerou um menor engajamento. Já o pico 12 é o segundo maior e ocorreu no jogo em que o Brasil foi desclassificado. O pico 13 correspondeu à final da competição e é o terceiro maior pico observado. É importante ressaltar que os picos no gráfico estão presentes nas linhas de tweets e retweets, ocorrendo sempre em conjunto e com relação a algum evento real. Isso indica como esses eventos impulsionaram o debate sobre a Copa do Mundo de 2022 no Twitter durante o período analisado.

⁴ <https://www.bbc.com/portuguese/internacional-63679803>

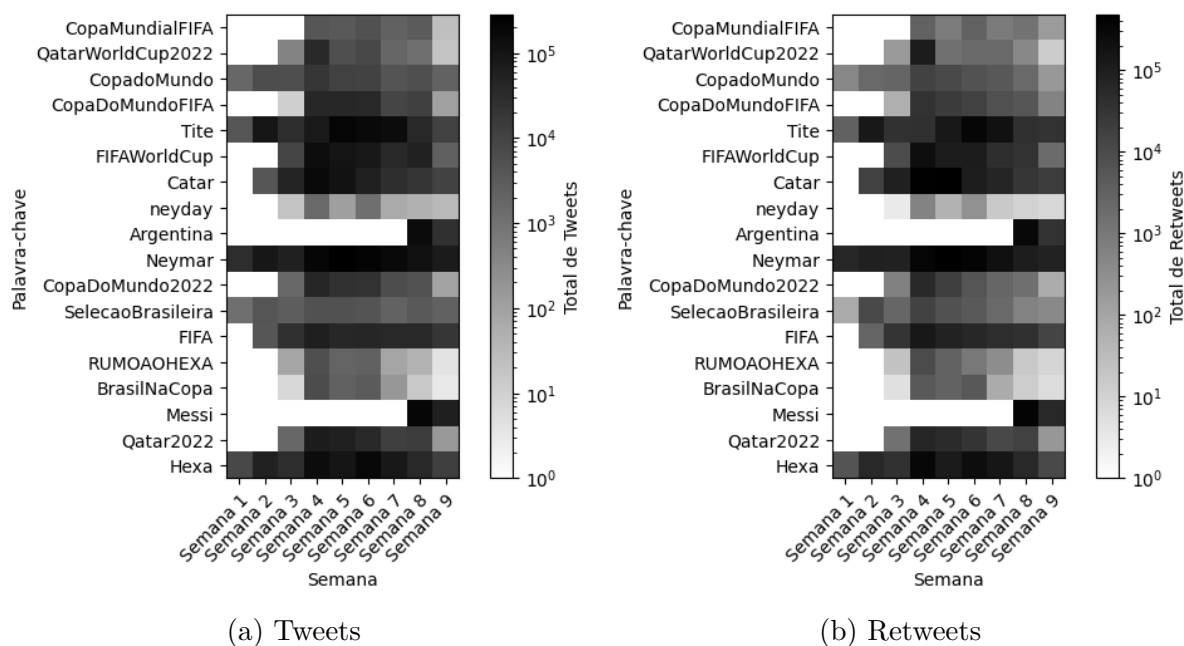
⁵ <https://www.metropoles.com/saude/entorse-entenda-o-que-e-lesao-sofrida-por-neymar-na-copa-do-m>

Figura 1 – Volume de tweets durante o período coletado



A seguir, foi feita uma análise detalhada dos tweets e retweets em três períodos de tempo abrangidos pelos dados coletados. A Figura 2 apresenta nuvens de palavras com os 100 termos mais frequentes em tweets e retweets durante a primeira, quinta e nona semana. A palavra *neymar* foi a mais comumente mencionada tanto nos tweets quanto nos retweets em todas as semanas. Além disso, pode-se observar que a palavra *brasil* foi bastante presente nas primeiras semanas, juntamente com a palavra *hexa*. No entanto, na nona semana, o termo *brasil*, que foi eliminado na semana 6 de maneira frustrante na decisão por pênaltis contra o Croácia, dá lugar a *argentina* que foi a equipe vencedora do torneio. Por conta disso, *messi*, nome de um dos melhores jogadores do mundo atualmente, também se destaca como uma palavra frequente. É interessante notar a presença de termos políticos, como *bolsonaro* e *lula*, nessas discussões relacionadas à Copa do Mundo. Isso se deve em parte ao fato de que o período de coleta ocorreu durante um ano eleitoral no Brasil. O nome do candidato que perdeu a eleição, que era o presidente em exercício na época, é proeminente na nuvem de retweets da quinta semana. Além disso, a utilização da camisa da Seleção Brasileira como forma de expressar opiniões políticas também contribuiu para a associação do seu nome nessas discussões.

Figura 3 – Popularidade das palavras-chave ao longo das semanas



4.3 RESULTADOS E ANÁLISES

Nesta seção, é detalhada e examinada uma série de análises conduzidas para compreender as diferentes perspectivas e discussões em torno da Copa do Mundo de 2022 no Twitter. Em cada análise, é descrita a metodologia utilizada e, em seguida, apresentada as principais conclusões obtidas.

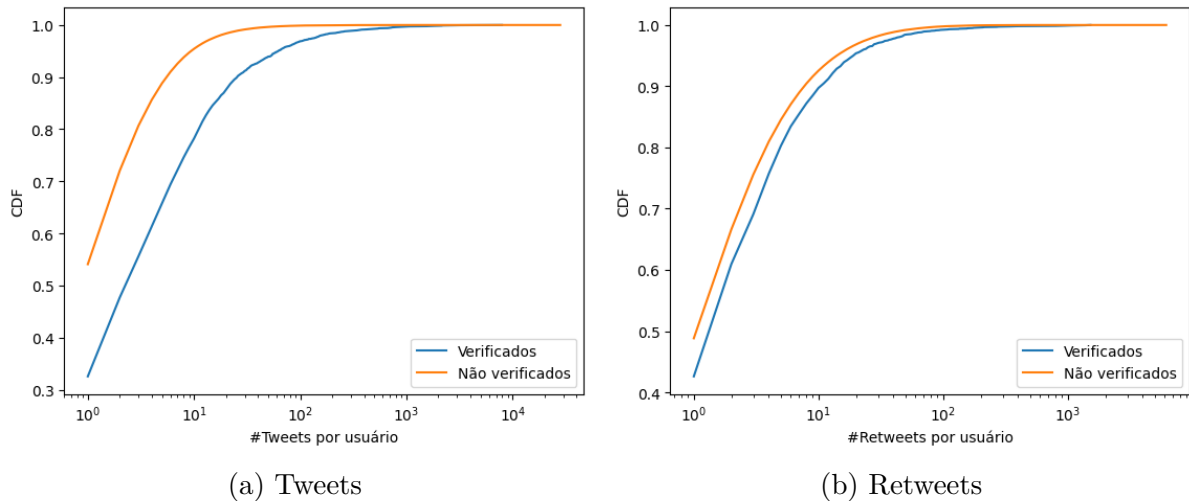
4.3.1 Perfil dos Indivíduos

A análise do perfil dos indivíduos tem como propósito descrever as características destes que participaram ativamente das discussões sobre a Copa do Mundo de 2022 no Twitter. Até o final de 2022, as pessoas que utilizavam a plataforma Twitter eram divididos em duas categorias de contas: verificadas e não verificadas. A verificação da conta pelo Twitter indicava que essas pessoas despertam um maior interesse público e tendiam a se envolver mais nas conversas relacionadas a eventos significativos para a sociedade (CHEN; LERMAN; FERRARA, 2020).

A Figura 4 ilustra as distribuições de probabilidade acumulada dos números de tweets e retweets feitos por pessoas com contas verificadas e não verificadas. Nos dados coletados as contas verificadas têm uma tendência maior a postar mais tweets e retweets. Cerca de 90% dos indivíduos com contas verificadas publicam até 30 tweets e 12 retweets, enquanto a mesma porcentagem com contas não verificadas publicam até 7 tweets e 9 retweets. O indivíduo mais ativo com uma conta verificada postou 7.930 tweets, enquanto o mais ativo com uma conta não verificada postou 28.084. Em relação aos retweets, verificou-se que as contas não verificadas mais ativas tendem a propagar mais informações, com um

máximo de 381.894 retweets para uma única pessoa, em comparação com apenas 3.841 retweets da mais ativa com conta verificada.

Figura 4 – Distribuições dos números de tweets e retweets por tipo de conta



Com o objetivo de analisar as palavras-chave mais comumente mencionadas pelas contas mais ativas, tanto verificadas quanto não verificadas, foram elaboradas as Tabelas 3 e 4. Essas tabelas mostram os cinco indivíduos mais ativos em cada categoria, juntamente com o número de seguidores e a frequência de ocorrência de cada palavra-chave em seus tweets e retweets. Observou-se que, embora essas contas mencionem palavras mais gerais, como *SelecaoBrasileiro*, *Qatar2022* e *Catar* com maior frequência, também foram encontradas menções ao técnico da seleção brasileira, conhecido como Tite, e ao jogador Neymar, que também é a palavra mais frequente nas nuvens de palavras. Entre as contas verificadas mais ativas em termos de número de tweets, notou-se que eram principalmente canais de notícias especializados em esportes. Além disso, foi observado que as contas mais ativas fizeram tweets e retweets relacionados a Argentina, que foi a campeã do torneio, e ao jogador Messi, que é o camisa 10 dessa seleção. É interessante destacar que, nas contas verificadas, os indivíduos mais ativos em termos de tweets possuíam um número muito maior de seguidores do que os mais ativos em termos de retweets, enquanto essa discrepância não foi observada nas contas não verificadas.

Tabela 3 – Top 5 contas mais ativas em # de tweets e retweets: contas verificadas

Tweets		
Conta	#Seguidores	Palavras-chave(#ocorrências)
1	858622	FIFAWorldCup(6013), CopaDoMundoFIFA(968), FIFA(751), Qatar2022(181), Tite(52), Neymar(42), Messi(22), Argentina(15), Hexa(15), CopadoMundo(10), CopaDoMundo2022(3), SelecaoBrasileira(2)
2	1518406	Qatar2022(2724), FIFAWorldCup(2686), Neymar(430), Tite(374), FIFA(187), Messi(115), Argentina(106), Hexa(66), Catar(19), CopadoMundo(18), SelecaoBrasileira(4), QatarWorldCup2022(2)
3	744417	Qatar2022(1002), CopadoMundo(762), CopaDoMundoFIFA(657), SelecaoBrasileira(161), Neymar(158), Tite(119), QatarWorldCup2022(102), CopaDoMundo2022(87), Argentina(63), FIFA(62), Messi(46), Hexa(44), FIFAWorldCup(10), Catar(9), CopaMundialFIFA(5), RUMOAOHEXA(1)
4	2701284	Qatar2022(1163), FIFAWorldCup(1108), Neymar(219), Tite(130), FIFA(71), Messi(36), Hexa(31), Argentina(30), CopadoMundo(17), Catar(13), CopaDoMundo2022(9), CopaDoMundoFIFA(8), CopaMundialFIFA(1), BrasilNaCopa(1)
5	268904	FIFAWorldCup(809), CopadoMundo(664), Qatar2022(571), SelecaoBrasileira(146), CopaDoMundoFIFA(141), Catar(137), Tite(64), Neymar(58), Hexa(23), Argentina(22), Messi(16), FIFA(12), CopaDoMundo2022(4), CopaMundialFIFA(2), QatarWorldCup2022(1)
Retweets		
Conta	#Seguidores	Palavras-chave(#ocorrências)
1	999	Qatar2022(3551), FIFAWorldCup(3190), Catar(341), Neymar(146), Tite(138), SelecaoBrasileira(88), Hexa(66), FIFA(65), Argentina(64), CopadoMundo(57), Messi(57)
2	998	Catar(2421), Neymar(160), Tite(122), FIFA(78), Argentina(38), Messi(31), Hexa(5), CopaDoMundo2022(2)
3	951	CopadoMundo(139), Catar(82), Qatar2022(63), SelecaoBrasileira(62), FIFAWorldCup(48), Neymar(42), FIFA(32), Tite(28), Argentina(18), CopaDoMundoFIFA(16), Hexa(14), Messi(11), CopaDoMundo2022(6), CopaMundialFIFA(3), QatarWorldCup2022(1)
4	999	Qatar2022(2563), FIFAWorldCup(2377), Tite(916), Neymar(837), Argentina(320), Messi(246), CopadoMundo(239), Catar(160), FIFA(156), Hexa(99), SelecaoBrasileira(5), QatarWorldCup2022(1)
5	970	CopadoMundo(263), Catar(32), Neymar(24), Argentina(14), Tite(13), SelecaoBrasileira(10), Messi(6), Hexa(4), FIFA(2), CopaDoMundoFIFA(1)

Tabela 4 – Top 5 contas mais ativas em # de tweets e retweets: contas não verificadas

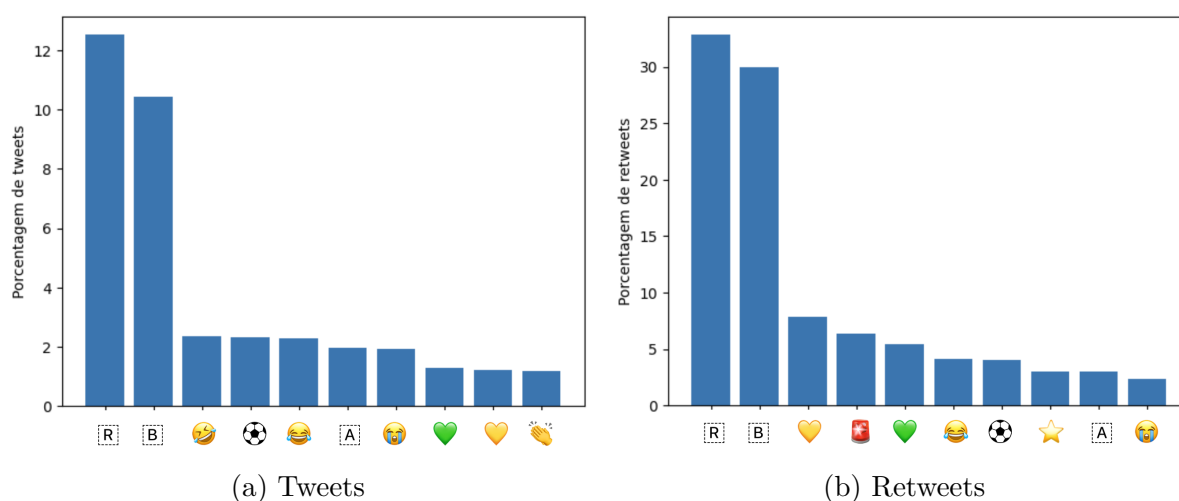
Tweets		
Conta	#Seguidores	Palavras-chave(#ocorrências)
1	956	SelecaoBrasileira(15300), CopadoMundo(15239), Catar(418), Neymar(274), FIFA(232), Argentina(223), Messi(192), Tite(188), Hexa(38), Qatar2022(6)
2	269617	Qatar2022(1966), FIFA(982), Catar(350), Neymar(206), Tite(136), Argentina(67), Messi(54), Hexa(28), FIFAWorldCup(6)
3	4463	FIFAWorldCup(1742), CopaDoMundoFIFA(716), CopaDoMundo2022(140), Qatar2022(124), CopaMundialFIFA(53), QatarWorldCup2022(29), Catar(26), FIFA(13), Hexa(6), Neymar(3), Messi(2), CopadoMundo(1), Tite(1), Argentina(1)
4	990	Qatar2022(1202), FIFAWorldCup(1158), Tite(29), Catar(24), FIFA(21), Hexa(15), Neymar(13), Argentina(6), Messi(4)
5	550	FIFAWorldCup(620), Qatar2022(619), QatarWorldCup2022(567), CopaDoMundoFIFA(487), CopaMundialFIFA(52), Tite(14), Neymar(8), FIFA(6), SelecaoBrasileira(3), CopaDoMundo2022(2), Hexa(2), Messi(1)
Retweets		
Conta	#Seguidores	Palavras-chave(#ocorrências)
1	999	Catar(130008), Neymar(78670), Tite(47070), FIFAWorldCup(37831), Hexa(37688), QatarWorldCup2022(28055), FIFA(13353), Argentina(8124), Messi(7069), CopaDoMundoFIFA(53)
2	9999	Neymar(122622), Hexa(48627), Tite(18782), Messi(7919), RUMOAOHEXA(3122), Catar(2918), FIFA(1071), FIFAWorldCup(243), Argentina(17), CopadoMundo(1), ney-day(1)
3	9999	Neymar(31057), Hexa(23850), SelecaoBrasileira(21896), Tite(17381), FIFA(10537), Messi(8141), Argentina(4092), CopadoMundo(670), Catar(15), Qatar2022(4)
4	9999	FIFAWorldCup(70694), QatarWorldCup2022(4906), Catar(2747), Hexa(2592), Qatar2022(1851), CopaDoMundoFIFA(1703), Tite(1531), Neymar(1300), CopaDoMundo2022(827), Messi(404), Argentina(290), FIFA(157), CopaMundialFIFA(88)
5	99974	Hexa(30962), Tite(8097), Neymar(5675), Argentina(4501), Messi(3940), FIFAWorldCup(531)

4.3.2 Análise dos Emojis

Os emojis são símbolos visuais usados nas redes sociais para complementar a comunicação por meio de texto. Eles representam emoções, ideias ou simbolismos. Segundo a análise dos dados, uma parte dos tweets (16,8%) e retweets (33,6%) continha pelo menos um emoji, identificados com o auxílio do pacote emoji⁶.

A Figura 5 apresenta os emojis mais utilizados entre os internautas que debatiam sobre a Copa do Mundo de 2022 no Twitter. Entre esses emojis que se destacaram, o emoji mais popular foi o do desenho da letra *R*, amplamente utilizado em tweets relacionados ao Brasil e à Argentina, representando as abreviações *BR* e *AR* para esses países. Os emojis de desenho das letras *B* e *A* também foram frequentemente usados em conjunto com o emoji do desenho da letra *R*. Outros dois emojis populares foram os corações nas cores verde e amarela, simbolizando o apoio à Seleção Brasileira no torneio. O emoji da bola de futebol também teve um uso significativo, o que era esperado dado o contexto da competição.









Figura 5 – Top-10 emojis mais frequentes em tweets e retweets



Além dos emojis mencionados anteriormente, é notável a frequência elevada dos emojis de choro, que podem expressar tanto tristeza quanto felicidade. O choro de tristeza foi amplamente utilizado em decorrência da derrota do Brasil no torneio e das lesões sofridas por alguns jogadores brasileiros, que resultaram em desfalques na equipe. Por outro lado, os emojis de choro de felicidade, popularmente conhecidos como *chorar de rir*, foram bastante utilizados em diferentes contextos. Em alguns momentos, foram empregados para comentar situações engraçadas, enquanto em outros foram usados para expressar deboche. Na Tabela 5, é possível encontrar exemplos do uso desses emojis, incluindo casos em que os emojis de choro de rir foram utilizados para ridicularizar críticas direcionadas ao jogador Neymar, que afirmavam que ele não era necessário para o time.

⁶ <https://pypi.org/project/emoji/>

Tabela 5 – Exemplos do uso de emojis

Emoji	Tweet
	QUEM NÃO TA ILUDIDO PRO HEXA É MALUCO
	Feliz com a vitória dos vizinhos argentinos. Grande jogo de Messi, que merecia muito, e Di Maria. Parabéns jogadores e comissão técnica da Argentina
	Neymar não faz falta não... Eu que faço
	Finalmente mês de novembro chegou o mês da CopadoMundo do Catar 2022, RumoAoHexa!
	ATUALIZAÇÃO SOBRE O NEYMAR: Neymar tem uma lesão recorrente no tornozelo e os médicos avaliam que esse pode ser um ponto favorável para seu retorno. Especialistas acreditam que sua recuperação pode ser mais fácil que a de Danilo! A maior preocupação com Neymar é o inchaço.
	Lionel Messi em Copas do Mundo: 5 participações, 26 jogos (recorde das Copas), 13 gols, 8 assistências, único a dar assistências em 5 Copas, único a marcar em todas as fases de mata-mata, 2x Bola de Ouro (2014 e 2022), 2 finais (2014 e 2022) e 1 título (2022)
	É hoje! Depois de muita espera, chegou o grande dia! Às 16h (de Brasília), a Seleção Brasileira estreia na Copa do Mundo FIFA Qatar 2022. Contamos com o seu apoio. Veste a Amarelinha e bora torcer por mais uma estrela! Vamos, Brasil!
	ArgenTRIna Que aula de futebol, que vitória mais linda. Parabéns Argentina, em especial ao Messi. Viva o futebol da América do Sul

Essas análises indicam que os emojis desempenharam um papel importante na comunicação dos internautas no Twitter durante as discussões sobre a Copa do Mundo de 2022, representando sentimentos, apoio, humor e ironia.

4.3.3 Análise de Sentimentos

Nesta seção, foi realizada uma análise dos sentimentos expressos nos tweets e retweets com base em diferentes palavras-chave. O mapa de calor apresentado na Figura 6 mostra essa análise através da pontuação contrastiva dos sentimentos, sendo calculada como a diferença entre a fração de tweets positivos e negativos.

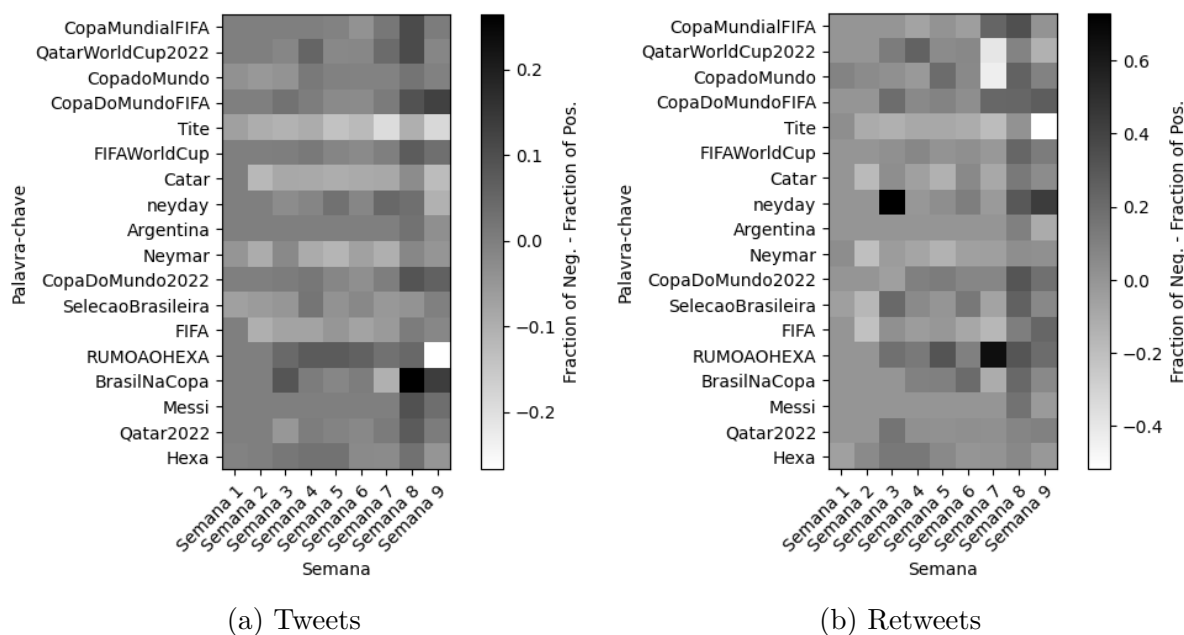
Na Figura 6a, podem ser observados os resultados para os tweets. Os posts que mencionam a palavra-chave *Tite*, referente ao técnico da Seleção Brasileira na competição, tendem a ser mais negativos (tons mais claros), principalmente nas últimas semanas da coleta dos dados. Como pode ser visto no tweet abaixo:

"#linhadepasse Tite errou muito quando não podia"

A palavra-chave *RUMOAOHEXA* teve uma maior fração de posts positivos na maioria das semanas, mas registrou uma maior fração de posts negativos na última semana. Isso provavelmente está relacionado ao fato de o Brasil ter perdido a chance de conquistar o sexto título de Copas do Mundo. Outras duas palavras-chave com uma fração ligeiramente negativa foram *Catar* e *Neymar*, o que está provavelmente ligado a críticas. Como mostrado respectivamente nos tweets a seguir:

"Queria que os dois times tivessem perdido e que o catar sumisse, quanta gente podre, credo que copa horrível"

Figura 6 – Evolução semanal da diferença entre a fração de sentimentos positivos e negativos



"E a má condição física do Neymar é ruim porque o time ficou mais previsível nos movimentos, já que é o único com liberdade dentro desse ataque mais posicional no ataque. E também não conseguiu grudar o time por causa da lesão, teve que dosar não correria."

Os tweets com maior fração de sentimento positivo tendem a ser mais para as últimas semanas, como observado nas palavras-chave *CopaMundialFIFA*, *QatarWorldCup2022*, *CopaDoMundoFIFA*, *FIFAWorldCup*, *CopaDoMundo2022*, *BrasilNaCopa*, *Messi* e *Qatar2022*. A maioria dessas palavras-chave está relacionada ao torneio de forma geral. Abaixo está um tweet que demonstra esse sentimento com essas palavras:

"Parabéns Argentina, Tricampeã Mundial de Futebol. Mereceu demais. #FIFAWorldCup #CopaMundialFIFA #ArgentinaVsFrance #Qatar2022"

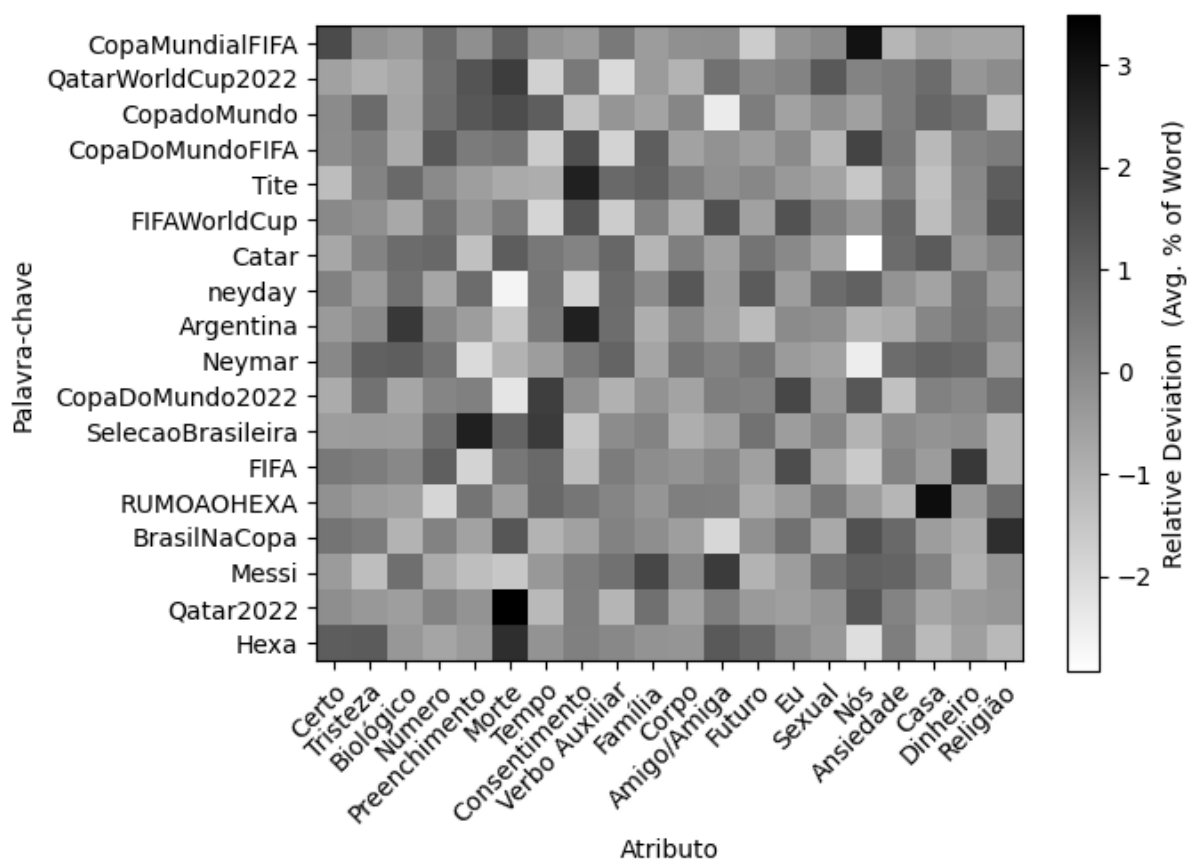
Além disso, foi analisada a pontuação contrastiva dos retweets, que podem ser considerados como uma medida da intensidade de difusão da informação nas redes sociais. A Figura 6b mostra que, em geral, os retweets seguem um padrão semelhante aos tweets. No entanto, na semana 7, as palavras-chave *QatarWorldCup2022* e *CopaDoMundo* registraram uma fração mais negativa nos retweets, enquanto nos tweets o registro foi mais neutro. Por outro lado, na última semana, a palavra-chave *RUMOAHEXA* apresentou um sentimento mais positivo nos retweets em comparação aos tweets.

Dessa forma, foi possível observar como o sentimento expresso pelas pessoas varia em relação a cada palavra-chave, tanto nos tweets quanto nos retweets, ao longo das semanas analisadas.

4.3.4 Análise Psicolinguística

Nesta seção, foram realizadas análises psicolinguísticas dos tweets e retweets relacionados ao debate sobre a Copa do Mundo de 2022. A Figura 7 mostra um mapa de calor com essa análise, considerando todos os tweets e retweets.

Figura 7 – Top LIWC atributos extraídos dos tweets e retweets coletados



Os resultados revelam diferenças nos atributos selecionados para as postagens relacionadas às palavras-chave mencionadas. Por exemplo, as postagens associadas a termos amplos relacionados ao torneio, como *CopaMundialFIFA*, *QatarWorldCup2022*, *CopaDoMundo* e *Qatar2022*, frequentemente fazem uso de palavras relacionadas à morte. Além disso, *CopaMundialFIFA* e *CopaDoMundoFIFA* têm postagens que utilizam com frequência palavras relacionadas à coletividade, representada por *nós* na figura. As palavras-chave *Tite* e *Argentina* estão mais associadas a palavras que transmitem a ideia de consentimento. Em relação às palavras-chave mais relacionadas ao Brasil na competição, como *SelecaoBrasileira*, *RUMOAOHEXA*, *BrasilNaCopa* e *Hexa*, foi observado que as postagens frequentemente empregam palavras relacionadas à ansiedade, coletividade, morte, religião, amizade, certeza e, principalmente, a palavra *casa* no caso de *RUMOAOHEXA*.

O atributo morte chama atenção no meio de um tema de entretenimento como a Copa do Mundo de 2022, nos dados coletados ele foi utilizado de algumas formas

diferentes. No contexto das palavras chave mais relacionadas a competição como um todo, como *CopaMundialFIFA*, *QatarWorldCup2022*, *CopaDoMundo* e *Qatar2022*, este atributo pode ser visto em tweets com expressões da língua portuguesa que não necessariamente remetem a morte, como pode ser visto abaixo:

"Mais um 0 x 0 na Copa, que morte horrível #Qatar2022 #FIFAWorldCup"

Outra maneira que trouxe este atributo nos tweets foi no caso de tweets que viralizaram e acabaram aparecendo muitas vezes através dos retweets, como o exemplo abaixo:

"@FIFAWorldCup quem é Argentina? Para o cego, é a escuridão. Para o faminto, é a fome. Para o sedento, é a sede. Para o morto, é a morte. Para o enfermo, é a piora. Para o prisioneiro, é a prisão. Para o solitário, é o desamparo. Para o viajante, é a desorientação. Para mim, é nada."

Além destas formas mostradas acima, nas palavras mais relacionadas ao Brasil e ao hexacampeonato, como *SelecaoBrasileira*, *RUMOAOHEXA*, *BrasilNaCopa* e *Hexa*, este atributo pode aparecer em tweets relacionados ao jogador Edson Arantes do Nascimento, mais conhecido como Pelé, que é considerado uns dos melhores jogadores da história e foi internado no período da competição. Ele faleceu logo após o final do evento, no dia 29/12/2022⁷. Abaixo está um exemplo de tweet utilizado desta forma.

"Não, o Pelé não pode morrer, ele tem que ver o hexa primeiro. Infelizmente a morte é para todos"

Em resumo, o LIWC se mostrou uma ferramenta útil para analisar o conteúdo do debate sobre a Copa do Mundo de 2022 no Twitter, fornecendo uma visão das narrativas dominantes no conjunto de dados analisados. Esta caracterização dos dados teve como objetivo a verificação de opiniões para facilitar a identificação de comunidades com opiniões semelhantes, analisadas no capítulo seguinte deste trabalho.

⁷ <https://ge.globo.com/pele/noticia/2022/12/29/morre-o-rei-pele-aos-82-anos.ghtml>

5 ANÁLISE E DETECÇÃO DE COMUNIDADES

O foco principal deste capítulo é analisar a existência de possíveis comunidades de internautas construídas a partir das interações ocorridas entre os mesmos. Neste sentido, este capítulo propõe um modelo em grafo para representar as interações entre eles. A partir da estrutura topológica encontrada, replicamos as análises realizadas no capítulo 5, considerando as maiores comunidades encontradas no grafo resultante. Sendo assim, estão disponíveis análises de perfil dos indivíduos, de emojis, de sentimentos, psicolinguística e de tópicos LDA. Em algumas das análises fez-se necessário que as informações fossem comparadas entre os dias escolhidos, nestes casos a comparação estará após a apresentação das informações desses dias na seção 5.6 Análise Comparativa.

5.1 TIPOS DE ANÁLISES REALIZADAS

Na etapa de análise e detecção de comunidades, dado o imenso volume de dados disponíveis, foram selecionados alguns dias específicos para a construção dessas redes, sendo o dia da cerimônia de abertura do evento, o dia da estreia da seleção brasileira, o dia da eliminação da seleção brasileira e o dia da partida final.

A rede gerada para rodar o algoritmo de detecção de comunidades foi uma rede baseada em retweets comuns entre pessoas. Estas redes foram representadas como grafos não direcionados, nos quais os vértices correspondem as pessoas e as arestas indicam que elas retweetaram o mesmo conteúdo. O peso das arestas representa a quantidade de retweets em comum entre esses indivíduos no conjunto de dados. Portanto, se o peso de uma aresta for, por exemplo, 10, isso significa que essas pessoas retweetaram 10 tweets idênticos, ou seja, com o mesmo ID, não apenas o mesmo texto.

A finalidade de representar o grafo dessa maneira foi criar uma rede de opinião, na qual a conexão entre os internautas é mais forte quanto mais informações compartilhadas eles têm em comum, como retweets do mesmo conteúdo. Nestas redes o algoritmo de detecção de comunidade de Louvain foi utilizado para obter as comunidades existentes neste grafo. Por conta de limitação computacional, este algoritmo de detecção de comunidades foi escolhido por ser eficiente (AIRES; NAKAMURA, 2017).

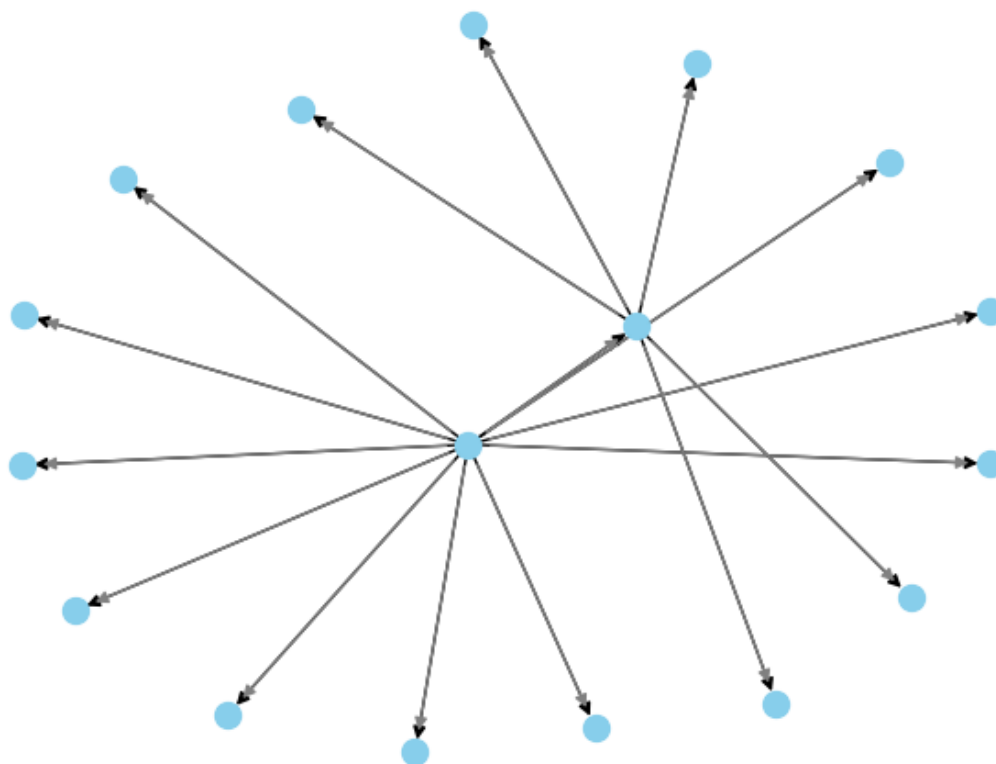
Após a detecção das comunidades, foram realizadas análises em relação às 5 maiores comunidades em termos de número de internautas. Isso incluiu a criação de nuvens de palavras conjugadas mais frequentes, análises de sentimentos, análises psicolinguísticas e análise dos indivíduos mais ativos e influentes.

A qualidade das comunidades obtidas pelo algoritmo de Louvain foi avaliada usando a métrica de modularidade, que indica o grau de conexão entre os elementos de uma comunidade. Essa métrica varia de -1 a 1 na implementação do algoritmo de Louvain utilizada. Nos dias selecionados, foram observadas modularidades próximas a 1, o que sugere uma detecção eficaz de comunidades, pois, nessa implementação, valores mais

próximos de 1 indicam uma detecção de alta qualidade.

Ainda na etapa de análise e detecção de comunidade foram feitas análises dos indivíduos mais influentes e mais ativos das comunidades, de nuvens de pares de palavras das comunidades, análise de sentimento das comunidades, análise psicolinguística e análise de tópicos. Na análise de indivíduos mais influentes e mais ativos foi necessário gerar uma rede de interação entre eles para cada comunidade, sendo esta os retweets, ou seja, o indivíduo que retweetou está ligado ao que fez o tweet. Estas redes foram representadas como grafos direcionados, nos quais os vértices correspondem aos internautas. A existência de uma ligação entre duas pessoas ocorre quando um deles posta um tweet e o outro realiza um retweet, que seria a interação nesse caso. A direção da ligação parte de quem fez o tweet e se dirige a quem interagiu com ele. Devido a essa abordagem, é comum observar estruturas de rede no formato de estrela, com vários nós conectados a apenas um nó central, como ilustrado na Figura 8. Infelizmente, por conta de limitação computacional, não foi possível ilustrar graficamente as redes geradas.

Figura 8 – Exemplo de grafo com formato estrela



Em cada uma das redes de interação foi realizada uma análise dos 5 indivíduos mais influentes com base no grau de saída, ou seja, aqueles que receberam o maior número de interações de outros indivíduos em seus tweets. Além disso, também foram examinados

os 5 indivíduos mais ativos com base no grau de entrada, representando aqueles que mais interagiram com os tweets de outros indivíduos.

As nuvens de palavras conjugadas têm como propósito trazer uma ideia geral sobre quais eram os assuntos mais discutidos, pois quanto mais recorrentes forem os pares de palavras, maiores serão eles na nuvem.

As análises de sentimento e psicolinguística foram feitas da mesma forma citada na caracterização dos dados gerais, porém na análise de sentimentos foram analisados apenas os retweets de cada dia, por conta da rede ter sido gerada a partir deles. Na análise psicolinguística desta etapa também existe uma diferença, ao invés da comparação ter sido feita em cima das palavras-chaves utilizadas para a extração dos dados, foi feita entre as 5 maiores comunidades em número de pessoas identificadas em cada dia.

A análise de tópicos foi feita utilizando uma implementação do algoritmo LDA em python¹. Foram analisados os retweets de cada uma das comunidades que foram trabalhadas. Antes de gerar a análise foi necessário verificar a quantidade de tópicos ideal de se trabalhar, isso foi feito verificando a coerência que cada quantidade de tópicos tinha. Por conta disso, algumas comunidades possuem menos tópicos nas análises.

5.2 DIA DA CERIMÔNIA DE ABERTURA DA COPA DO MUNDO DE 2022

A rede gerada para o dia da cerimônia de abertura contava com 164.281 internautas e revelou a existência de 2.151 comunidades. A Tabela 6 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades.

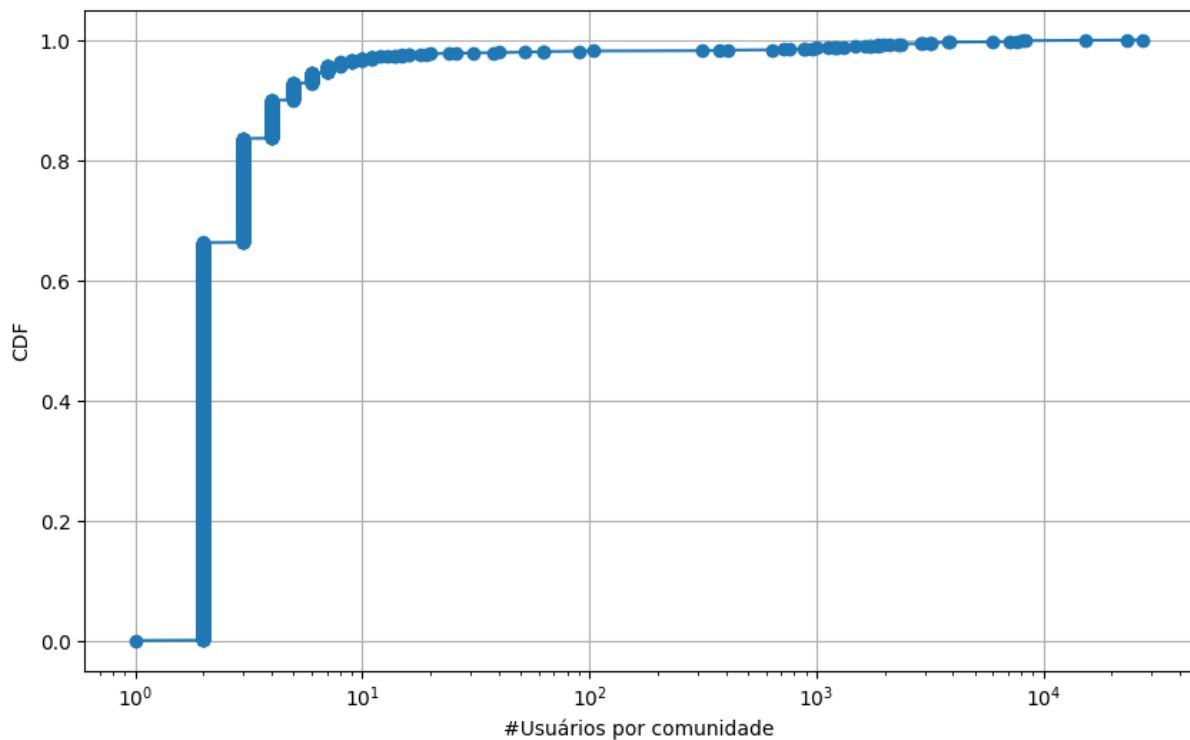
Tabela 6 – Informações gerais sobre os grafos do dia da cerimônia de abertura

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	164.281	12.055.873	146,77	-
Comunidade 1	27.350	2.493.228	182,32	7
Comunidade 2	23.425	770.294	65,77	9
Comunidade 3	15.253	592.159	77,64	10
Comunidade 4	8.278	261.549	63,191	9
Comunidade 5	8.077	243.416	60,27	10

As comunidades encontradas apresentaram uma modularidade de 0,70. A Figura 9 apresenta a distribuição do número de indivíduos por comunidade neste dia, essa distribuição está sendo comparada com os outros dias na seção 5.6 Análise Comparativa.

¹ <https://pypi.org/project/gensim/>

Figura 9 – Distribuição de número de indivíduos por comunidade nos dia da cerimônia de abertura



A Tabela 7 fornece o total de internautas e a quantidade de retweets em cada comunidade, além dos valores totais para este dia.

Tabela 7 – Número de internautas e retweets por comunidade no dia da cerimônia de abertura

Comunidade	Número de internautas	Número de retweets
1	27.350	84.117
2	23.425	56.747
3	15.253	40.587
4	8.278	15.991
5	8.077	16.794
Todas	164.281	388.540

Perfil dos Indivíduos

Na Tabela 8 está disponível o número de tweets e seguidores de cada um dos 5 indivíduos mais influentes de cada uma das comunidades deste dia, além de mostrar se é ou não verificado. A Tabela 9 possui o mesmo formato, porém mostra os 5 indivíduos mais influentes de cada uma das comunidades.

Tabela 8 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da cerimônia de abertura

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	1798	177138	Não	802	317	Não	728	2012	Não	278	155	Não	14	70351	Não
2	1650	263704	Não	179	15324	Não	652	18	Não	42	823285	Não	13	314545	Sim
3	1550	102717	Não	147	401	Não	423	715	Não	2	1006	Não	9	16	Não
4	891	5920	Não	52	61	Não	392	716	Não	1	119	Não	4	2726	Não
5	889	185176	Não	51	311	Não	342	911	Não	1	561	Não	3	7346	Não

Tabela 9 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da cerimônia de abertura

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	65	71	Não	80	210	Não	67	242	Não	55	854	Não	36	1100	Não
2	59	153	Não	52	79	Não	67	96	Não	40	41	Não	32	746	Não
3	50	341	Não	41	76	Não	60	4060	Não	36	285	Não	30	2582	Não
4	48	20091	Não	41	55	Não	59	413	Não	35	444	Não	28	424	Não
5	43	7	Não	40	6	Não	57	850	Não	34	3421	Não	27	13	Não

Nuvens dos 50 Pares de Palavras mais Frequentes

Na Figura 10 estão disponíveis as nuvens dos 50 pares de palavras mais frequentes. Na comunidade 1, a maior comunidade em número de internautas, os pares de palavras que mais apareceram foram *catar qatarworldcup2022*, *mundo qatarworldcup2022* e *jungkook bts*. Essas parecem estar mais ligadas ao início do evento de modo geral. Jungkook é um integrante da banda BTS, que se apresentou na cerimônia de abertura². Além disso as palavras *catar* e *qatarworldcup2022* fazem parte das palavras chaves utilizadas na coleta dos dados. As comunidades 2, 4 e 5 destacaram pares de palavras semelhantes, apareceram em todas *país anfitrião*, *direitos humanos* e *primeiro país*. Discussões sobre os direitos humanos no Catar, país sede da Copa do Mundo de 2022, foram levantadas no período do início do evento³, o que pode ter gerado tanta recorrência dos pares de palavras citados. A nuvem gerada para a comunidade 3 está diferente das demais, ela está com pares de palavras de tamanho muito uniforme e com uma mistura das que já apareceram nas outras nuvens.

² <https://ge.globo.com/futebol/copa-do-mundo/noticia/2022/11/20/>

[bts-na-copa-do-mundo-do-catar-jungkook-na-abertura-agita-fas-de-k-pop.ghtml](https://ge.globo.com/futebol/copa-do-mundo/noticia/2022/11/20/bts-na-copa-do-mundo-do-catar-jungkook-na-abertura-agita-fas-de-k-pop.ghtml)

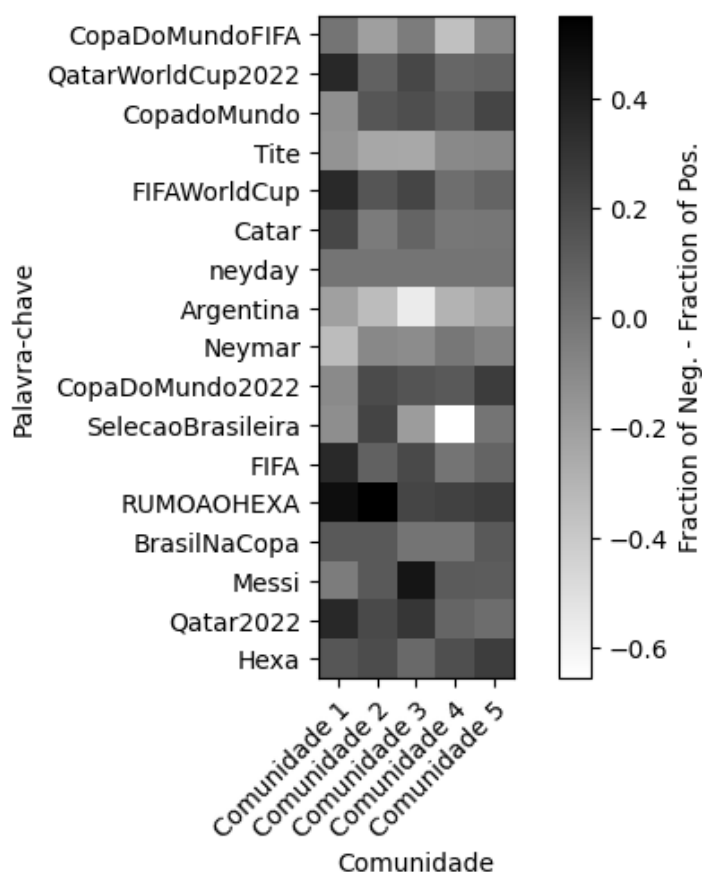
³ <https://www.cnnbrasil.com.br/esportes/copa-do-mundo-entenda-as-denuncias-sobre-direitos-humanos>

Argentina estão com sentimento mais negativo em todas as cinco comunidades por conta da rivalidade existente entre o Brasil e a Argentina. A seguir está um exemplo de tweet representando esse cenário:

"BORA BRAAAAASIL a meta é: Ser hexa Devolver o 7x1 pra Alemanha
Produzir vários memes Eliminar a Argentina #QatarWorldCup2022"

A palavra *Neymar*, que representa o jogador número 10 da seleção brasileira, está negativo apenas na comunidade 1 e *SelecaoBrasileira* apenas na comunidade 4. Um ponto interessante de se destacar é que a palavra-chave *Messi*, que é jogador número 10 da seleção argentina, está com sentimentos mais positivos nas comunidades, mesmo com a *Argentina* aparecendo com sentimentos negativos.

Figura 11 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da cerimônia de abertura



Análise Psicolinguísticas

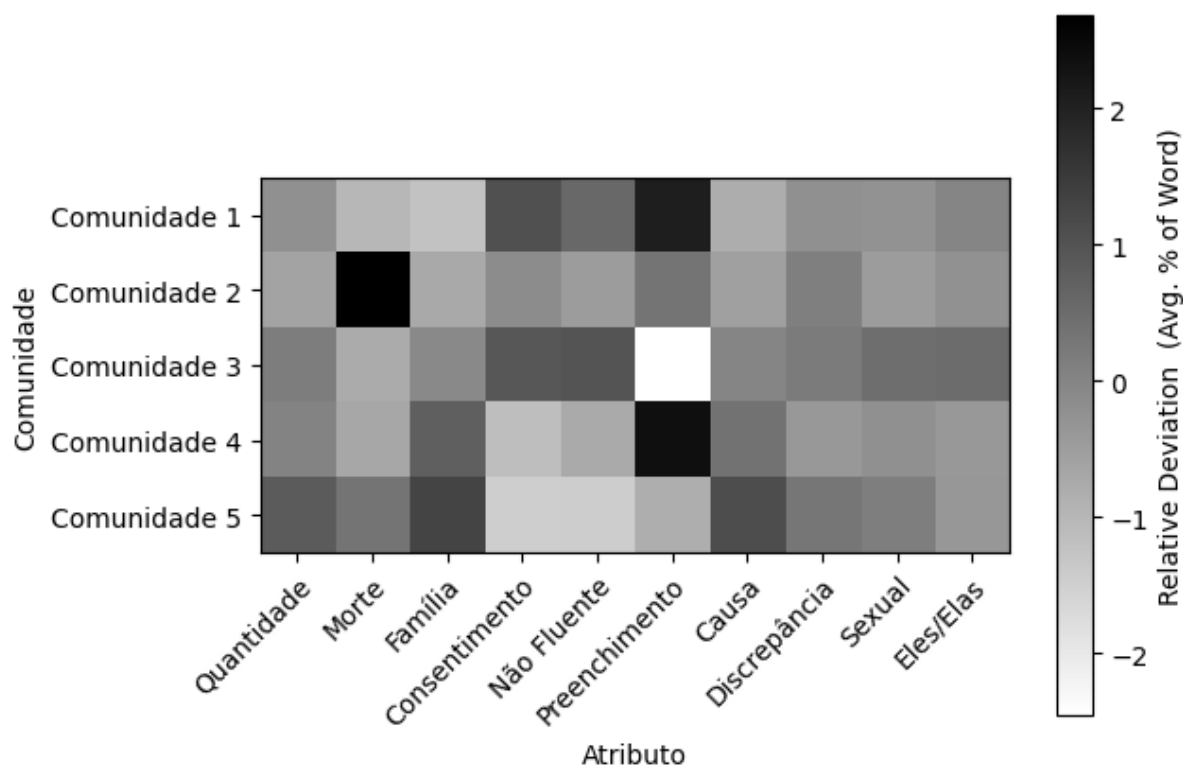
Na Figura 12 pode ser vista a análise psicolinguística. A comunidade 1, que é

a maior em número de internautas, utiliza frequentemente palavras de preenchimento⁴, bem como palavras relacionadas a consentimento e não fluência⁵ em seus retweets. A comunidade 2 se destaca pelo uso de palavras associadas à morte, muito por conta de tweets como esse:

"O Catar proibir cerveja pegou + no Brasil do que eles humilharem mulheres, terem atitudes racistas e pena de morte p/ população LBGBTQIA+, escravizarem migrantes, dentre outras atrocidades Alguns no Brasil precisam urgente de um banho de civilidade, respeito e amor no coração..."

A comunidade 3 está relacionada a palavras de consentimento e não fluência. A comunidade 4, por sua vez, concentra seus retweets em palavras de preenchimento. Por fim, a comunidade 5 tem uma conexão mais forte com tópicos relacionados à família e causas.

Figura 12 – Análise de psicolinguística das top 5 maiores comunidades do dia da cerimônia de abertura



Análise de Tópicos LDA

⁴ O atributo *preenchimento* (filler em inglês) envolve palavras ou expressões de preenchimento, como "uh", "é", "bem", que são usadas na fala para ganhar tempo ou preencher lacunas na conversa.

⁵ O atributo *não fluência* (nonfl em inglês) pode estar relacionada a palavras ou expressões que indicam falta de fluência na comunicação, como hesitações ou repetições.

A Tabela 10 mostra a análise de tópicos para a comunidade 1 do dia de abertura do evento. Nela foram identificados 10 tópicos, onde podemos notar que as discussões estão girando em torno da apresentação da banda BTS. O nome do cantor Jeon Jung-Kook, que se apresentou na abertura do evento como membro desta banda⁶, está aparecendo em alguns tópicos desta comunidade, assim como o nome da banda.

Tabela 10 – Análise de tópicos LDA para a comunidade 1 do dia da cerimônia de abertura

Tópico 1	jungkook	apresentou	copa	abertura	dreamers
Tópico 2	mundo	copa	abertura	bts	jungkook
Tópico 3	pop	jungkook	abertura	copadomundo2022	mundo
Tópico 4	primeiro	história	torna	ato	performar
Tópico 5	globonacopa	brasil	feito	copadomundo2022	rumoaohexa
Tópico 6	qatarworldcup2022	famoso	família	qatar2022	música
Tópico 7	imagina	orgulho	jeon	ponto	relevante
Tópico 8	sempre	qatarworldcup2022	sul	feliz	alguém
Tópico 9	estádio	simplesmente	ole	clima	música
Tópico 10	seleção	jungkook	coreia	pitico	visitando

Na tabela 11 está a análise para a comunidade 2, onde foram identificados 8 tópicos. Estes tópicos parecem mostrar algumas discussões diferentes. Os tópicos 1, 4, 6 e 8 parecem ter discussões em torno do Brasil no torneio. Nos tópicos 3 e 7 as discussões tendem a ser mais sobre assuntos do evento em geral. Além disso, nos tópicos 2 e 5 apareceram discussões em torno de direitos humanos e diversidades, a seguir estão tweets deste dia que exemplificam como foram tratados esses assuntos:

"HISTÓRICO: Catar se tornou o primeiro país anfitrião a ser derrotado em uma estreia de Copa do Mundo. Eles não tem direitos humanos e também não tem vitória. Parabéns, Catar! #FIFAWorldCup #QatarWorldCup2022"

"'Que beleza juntar todas essas diversidades aqui' #QatarWorldCup2022"

O tweet que fala a respeito de diversidades foi uma frase falada no discurso de abertura do evento⁷.

Tabela 11 – Análise de tópicos LDA para a comunidade 2 do dia da cerimônia de abertura

Tópico 1	brasil	hexa	fifaworldcup	fuleco	copa
Tópico 2	qatarworldcup2022	abertura	catar	direitos	humanos
Tópico 3	copa	qatarworldcup2022	mundo	abertura	waka
Tópico 4	qatarworldcup2022	brasil	equador	qatar2022	copa
Tópico 5	qatarworldcup2022	qatar	sempre	beleza	diversidades
Tópico 6	todas	tite	durante	copas	fifa
Tópico 7	qatarworldcup2022	partir	treze	copa	catar
Tópico 8	juntar	importa	tão	nesse	tite

⁶ <https://ge.globo.com/futebol/copa-do-mundo/noticia/2022/11/20/bts-na-copa-do-mundo-do-catar-jungkook-na-abertura-agita-fas-de-k-pop.ghtml>

⁷ <https://oglobo.globo.com/esportes/catar-2022/noticia/2022/11/que-beleza-juntar-essa-diversidade-toda-diz-emir-do-catar-em-discurso-na-abertura-da-copa.ghtml>

A Tabela 12 ilustra a análise de tópicos da comunidade 3, mostrando seus 10 tópicos identificados. Assim como nas outras comunidades, nesta também abordam assuntos gerais a respeito do torneio como um todo, além de falar sobre o Brasil na competição e falar sobre direitos humanos na abertura. Um tema diferente apareceu no tópico 5, é falado sobre a música Waka-waka da cantora Shakira que foi a principal música da Copa do Mundo de 2010⁸. Abaixo pode ser visto um tweet fala a respeito disso:

"a melhor música da copa do mundo continua sendo waka-waka de shakira
#QatarWorldCup2022 #CopaDoMundo2022 #shakira"

O tópico 10 também trouxe um tema novo, nele podemos ver a palavra *deus*, que nesse dia foi usada para falar da participação do ator Morgan Freeman na cerimônia de abertura⁹. Ele é popularmente conhecido como Deus por conta da atuação no filme *Todo Poderoso*, onde representava esse personagem¹⁰. No tweet abaixo podemos ver um exemplo de como o assunto foi citado:

"Simplesmente deus na #CopaDoMundo2022 #QatarWorldCup2022"

Tabela 12 – Análise de tópicos LDA para a comunidade 3 do dia da cerimônia de abertura

Tópico 1	brasil	qatarworldcup2022	fifaworldcup	começa	copa
Tópico 2	qatarworldcup2022	jungkook	abertura	copa	copadomundo2022
Tópico 3	qatarworldcup2022	clima	nesse	acordar	acontece
Tópico 4	abertura	qatarworldcup2022	cerimônia	mundo	copa
Tópico 5	waka	qatarworldcup2022	shakira	copa	acordei
Tópico 6	qatarworldcup2022	abertura	humanos	direitos	fingindo
Tópico 7	tocando	melhor	feito	juntar	rico
Tópico 8	qatarworldcup2022	sempre	argentina	brasil	seleção
Tópico 9	qatarworldcup2022	futebol	mostratuaforça	aparecendo	selecaodosamba
Tópico 10	copa	deus	qatarworldcuo2022	contratou	nada

Nas Tabelas 13 e 14 estão disponíveis, respectivamente, as análises das comunidades 4 e 5. Em ambas as comunidades foram identificados 2 tópicos. Esses tópicos estão abordando temas gerais a respeito da abertura da copa, início dos jogos e participação do Brasil no evento.

Tabela 13 – Análise de tópicos LDA para a comunidade 4 do dia da cerimônia de abertura

Tópico 1	catar	copa	brasil	equador	qatarworldcup2022
Tópico 2	catar	qatarworldcup2022	mundo	copa	país

⁸ <https://g1.globo.com/mundo/copa-do-catar/noticia/2022/11/19/a-historia-por-tras-de-waka-waka-o-maior-hit-das-copas-shakira-copiou-musica-camaronesa.ghhtml>

⁹ <https://ge.globo.com/futebol/copa-do-mundo/noticia/2022/11/20/festa-marca-a-abertura-da-copa-do-mundo-do-catar.ghhtml>

¹⁰ <https://www.adorocinema.com/filmes/filme-43219/>

Tabela 14 – Análise de tópicos LDA para a comunidade 5 do dia da cerimônia de abertura

Tópico 1	qatarworldcup2022	catar	copa	mundo	primeiro
Tópico 2	abertura	copa	brasil	qatarworldcup2022	catar

5.3 DIA DA ESTREIA DA SELEÇÃO BRASILEIRA NA COMPETIÇÃO

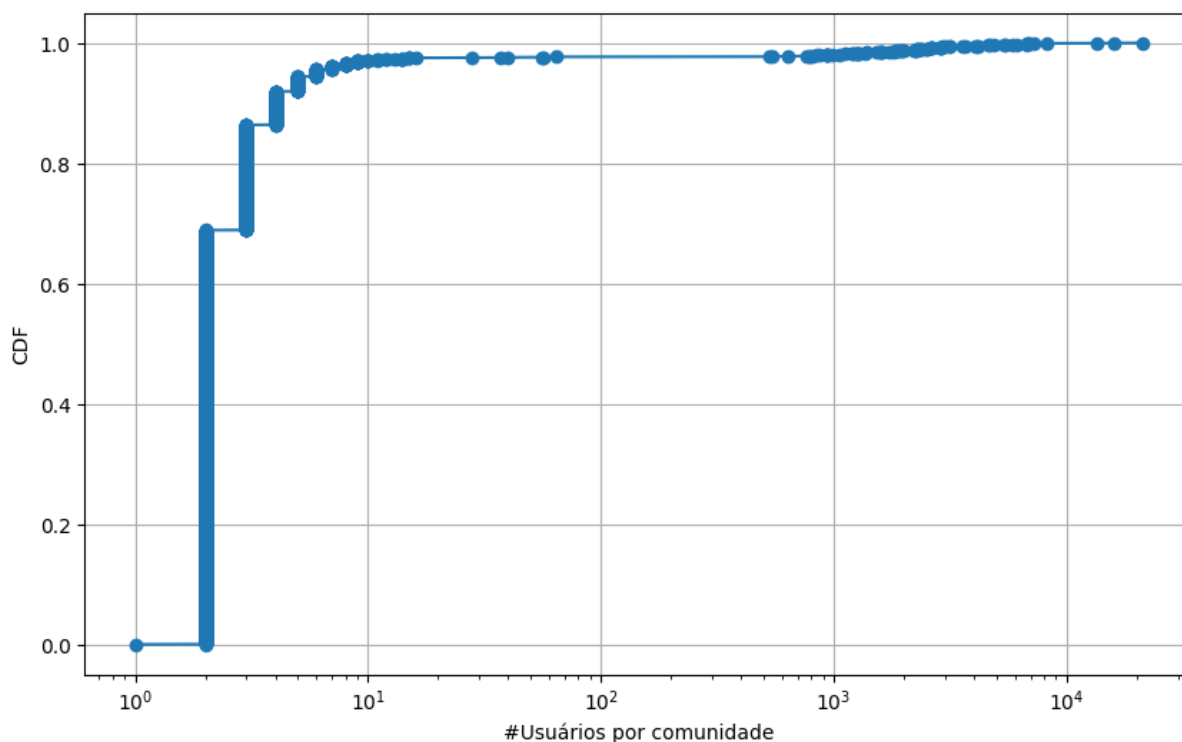
A rede gerada para o dia da estreia da seleção brasileira na competição incluía 266.622 internautas e resultou em 3.527 comunidades. A Tabela 15 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades.

Tabela 15 – Informações gerais sobre os grafos do dia da estreia da seleção brasileira

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	266.622	23.252.970	174,43	-
Comunidade 1	21.209	506.842	47,79	12
Comunidade 2	15.881	1.647.442	207,47	8
Comunidade 3	13.445	1.336.024	198,74	8
Comunidade 4	8.204	496.100	120,94	8
Comunidade 5	7.281	342.204	93	11

As comunidades encontradas apresentaram uma modularidade de 0,72. A Figura 13 apresenta a distribuição do número de indivíduos por comunidade neste dia, essa distribuição está sendo comparada com os outros dias na seção 5.6 Análise Comparativa.

Figura 13 – Distribuição de número de indivíduos por comunidade nos dia da estreia de seleção brasileira



A Tabela 16 fornece o total de internautas e a quantidade de retweets em cada comunidade, além dos valores totais para este dia.

Tabela 16 – Número de internautas e retweets por comunidade no dia da estreia da seleção brasileira

Comunidade	Número de internautas	Número de retweets
1	21.209	41.996
2	15.881	39.491
3	13.445	20.883
4	8.204	18.699
5	7.281	13.632
Todas	266.622	554.667

Perfil dos Indivíduos

Na Tabela 17 está disponível o número de tweets e seguidores de cada um dos 5 indivíduos mais influentes de cada uma das comunidades deste dia, além de mostrar se é ou não verificado. A Tabela 18 possui o mesmo formato, porém mostra os 5 indivíduos mais influentes de cada uma das comunidades.

Tabela 17 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da estreia da seleção brasileira

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	3149	169511	Não	360	538527	Não	215	2430534	Sim	401	448258	Sim	80	3947	Não
2	249	6238	Não	83	177742	Não	130	2167	Não	18	638	Não	8	110187	Sim
3	78	205125	Não	72	7044	Não	26	11436	Não	15	7522580	Sim	1	240	Não
4	70	73088	Não	48	18675	Não	25	19871	Não	4	1251	Não	1	643	Não
5	66	1875	Não	39	3011	Não	12	17123	Não	3	10321	Não	1	1024	Não

Tabela 18 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da estreia da seleção brasileira

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	47	113	Não	61	1600	Não	60	881	Não	62	202	Não	29	1281	Não
2	41	2023	Não	40	2170	Não	20	76	Não	38	290	Não	26	655	Não
3	35	684	Não	29	264	Não	18	5749	Não	28	35	Não	24	1355	Não
4	33	113	Não	29	114	Não	17	2819	Não	28	3416	Não	21	272	Não
5	29	33	Não	29	1222	Não	14	612	Não	27	384	Não	20	4598	Não

Nuvens dos 50 Pares de Palavras mais Frequentes

As nuvens dos 50 pares de palavras mais frequentes estão disponíveis na Figura 14. Na comunidade 1 aparecem frequentemente os pares de palavras *possível lesão*, *neymar seguimos*, *neymar merece*, *bizarro irmão*, *seguimos juntos* e *posicionamento político*. Esses pares de palavras provavelmente tem uma relação com uma lesão que o jogador Neymar sofreu no jogo de estreia¹¹, mas em especial o *posicionamento político* deve-se ao fato

¹¹ <https://www.metropoles.com/saude/entorse-entenda-o-que-e-lesao-sofrida-por-neymar-na-copa-do-m>

dele ter apoiado Jair Messias Bolsonaro, um dos candidatos a presidência do Brasil neste mesmo período da Copa do Mundo¹². As comunidades 2, 4 e 5 estão tratando mais assuntos relacionados a vitória do Brasil no jogo, os pares de palavras mais frequentes foram *pombo richarlison*, *richarlison abre*, *brasil vence*, *richarlison marca*, *sérvia fifaworldcup* e *brasil rumo*. O nome do jogador Richarlison foi muito citado por ele ter sido quem marcou os gols pela seleção brasileira¹³. A nuvem da comunidade 3 faz referência especialmente a propagação do tweet:

"Só pra lembrar.. o 1º a cair na luta pela democracia, e ter livre arbítrio e liberdade, já tentava nos alertar, foi um herói policial militar, a #CopaDoMundo2022 não vai trazer nossa liberdade, 'Meu Povo Padeceu por falta de Conhecimento' @JaspionRedPiLLL"

Os principais pares de palavras são parte dele, sendo elas *herói policial*, *povo padeceu*, *livre arbítrio* e *policial militar*. Infelizmente não foram encontradas mais informações para o esclarecimento desse tweet.

¹² <https://oglobo.globo.com/esportes/noticia/2022/10/ney-mar-diz-que-vai-comemorar-primeiro-gol-na-gh.html>

¹³ <https://www.cnnbrasil.com.br/esportes/com-dois-gols-de-richarlison-brasil-vence-a-servia-na-es>

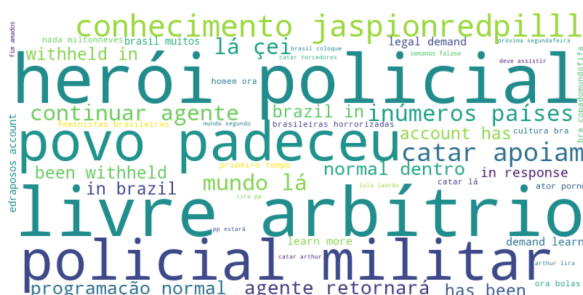
Figura 14 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da estreia da seleção brasileira



(a) Comunidade 1



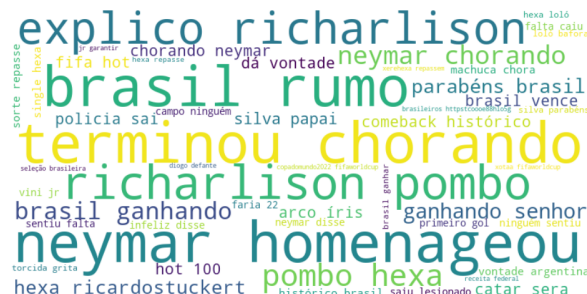
(b) Comunidade 2



(c) Comunidade 3



(d) Comunidade 4



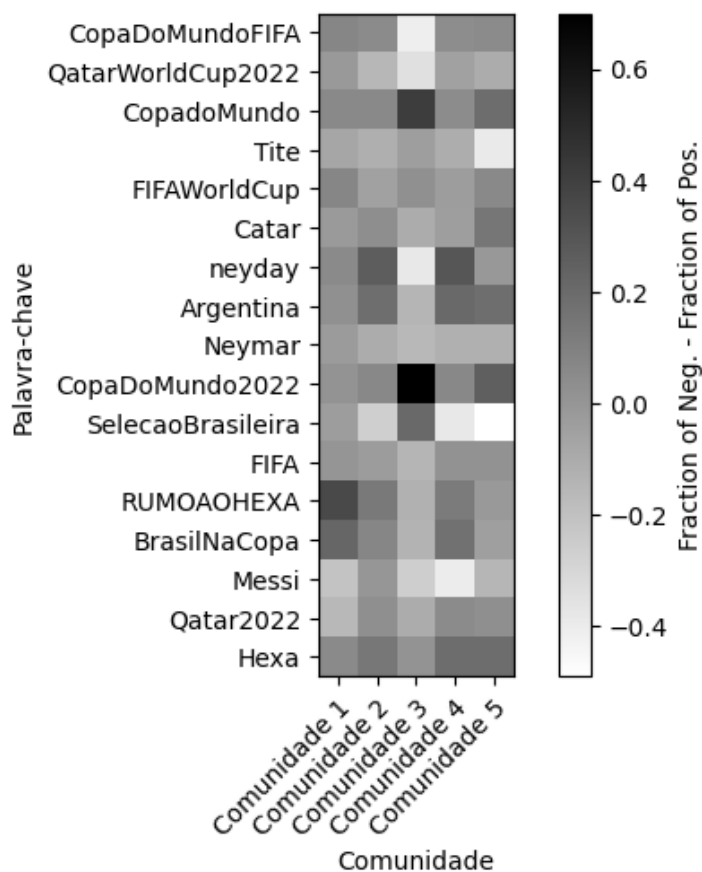
(e) Comunidade 5

Análise de Sentimentos

A Figura 15 apresenta a análise de sentimentos. Dentre os dias analisados, esse foi o que apresentou mais sentimentos negativos. A comunidade 1, a maior em número de internautas, apresenta sentimentos positivos para as palavras-chave *RUMOAOHEXA* e *BrasilNaCopa*, demonstrando um ânimo relacionado a estreia da seleção brasileira. As comunidades 2, 4 e 5 tiveram sentimento negativo relacionado às palavras *SelecaoBrasileira* e *Tite*, o que pode significar uma possível insatisfação a respeito do jogo de estreia. A comunidade 3 obteve sentimento mais negativo para palavras-chave relacionadas ao Brasil na copa, como *RUMOAOHEXA*, *BrasilNaCopa*, *Neymar* e *neyday*. Além de apresentar sentimentos bastante positivos e negativos relacionados ao evento de modo geral, com

CopaDoMundoFIFA e *QatarWorldCup2022* com sentimentos muito negativos e *CopaDoMundo2022* com sentimento muito positivo. Diferente do dia da cerimônia de abertura, a palavra *Messi* apareceu com sentimento mais negativo nessas comunidades.

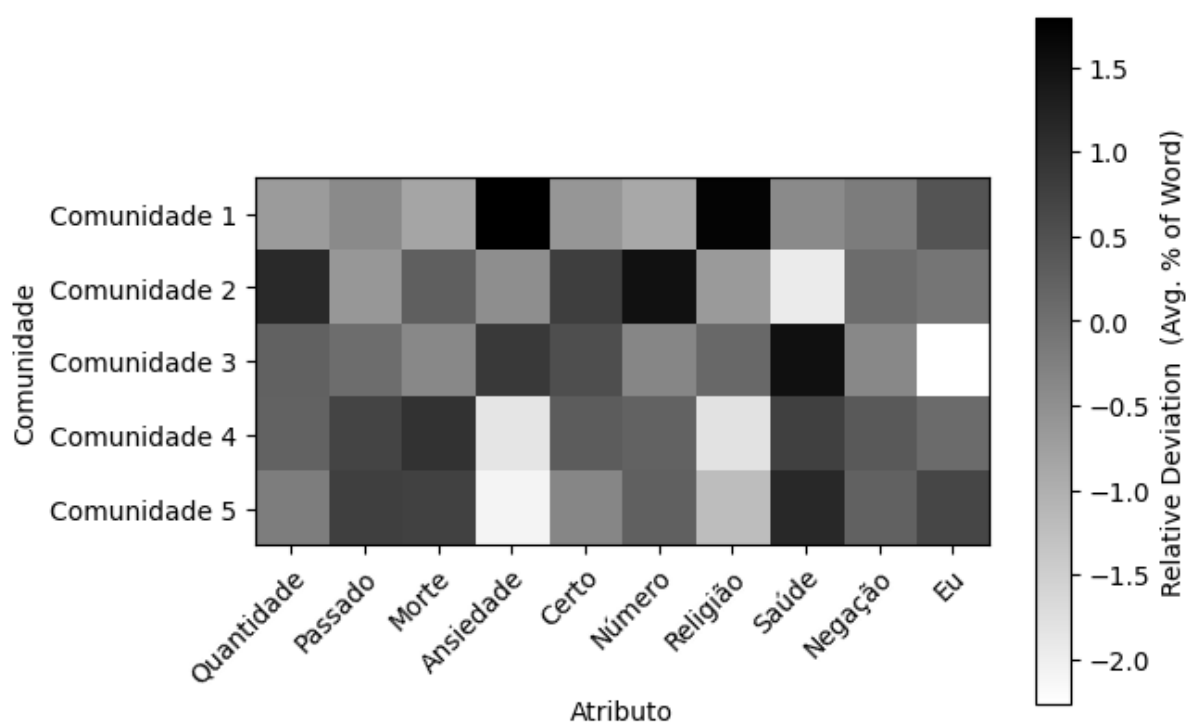
Figura 15 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da estreia da seleção brasileira



Análise Psicolinguísticas

A análise psicolinguística pode ser visualizada na Figura 16. Nessa análise, pode-se notar que na comunidade 1 os retweets estão mais relacionados à ansiedade e religião. Na comunidade 2 predominam palavras relacionadas a quantidade e número. A comunidade 3 está muito associada à saúde e à ansiedade. As comunidades 4 e 5 se destacam por tópicos relacionados ao passado, saúde e morte.

Figura 16 – Análise de psicolinguística das top 5 maiores comunidades do dia da estreia da seleção brasileira



Análise de Tópicos LDA

Na tabela 19 está a análise de tópicos para a comunidade 1 do dia da abertura do evento. Nela foram identificados 10 tópicos e o assunto principal parece ser o jogo da seleção brasileira e a lesão do jogador Neymar. Os tópicos 3 e 4 discutem a escolha da titularidade de alguns jogadores, como pode ser visto no tweet abaixo:

"Não entra na minha cabeça esse Raphinha titular. Tanto GABIGOL quanto Everton Ribeiro pra mim são melhores que ele. #CopaDoMundoFIFA "

Os tópicos 1, 2, 5, 6, 8, 9 e 10 parecem falar sobre a lesão do Neymar, os tweets abaixo exemplifica como algumas pessoas discutiam este assunto:

"Tem gente comemorando a lesão do Neymar por causa do posicionamento político dele. Que bizarro, irmão."

"esse povo que se diz torcedor da seleção e tá rindo da possível lesão tô neymar merece todo desprezo possível que nojo"

Tabela 19 – Análise de tópicos LDA para a comunidade 1 do dia da estreia da seleção brasileira

Tópico 1	neymar	lesão	jogar	comemorando	deus
Tópico 2	tite	neymar	copa	seleçãobrasileira	começar
Tópico 3	time	mundo	copa	titular	gabigol
Tópico 4	raphinha	jr	vini	gols	minutos
Tópico 5	qatar2022	brasil	tornozelo	esportudonacopa	grêmio
Tópico 6	favor	possível	nojo	desprezo	rindo
Tópico 7	hexa	catar	caminhada	partiu	rapaziada
Tópico 8	neymar	participou	bola	técnica	descobriram
Tópico 9	tirou	deve	xingando	neyday	bêbada
Tópico 10	político	deixou	mano	deixar	bagre

A Tabela 20 mostra a análise da comunidade 2, onde foram identificados 6 tópicos. O primeiro aborda o desempenho do jogador Richarlison, que fez dois gols neste jogo de estreia¹⁴. O tópico 2 fala sobre a lesão do jogador Neymar, conforme já foi comentado na análise da comunidade 1. Os tópicos 3, 5 e 6 possuem muitos tweets com relação a torcida pelo Brasil na competição, conforme pode ser visto no seguinte tweet:

"HOJE É DIA DE FESTA NO BRASIL Daqui a pouco a seleção brasileira entra em campo para seu primeiro jogo na #CopaDoMundo2022, e estamos na torcida pelo hexa, com muita festa!"

No tópico 5 aparece *bolsonaro*, este é o sobrenome do presidente do Brasil até 2022. Jair Bolsonaro também foi um dos candidatos às eleições presidenciais que ocorreram em 2022, pouco antes da Copa do Mundo. A camisa da seleção brasileira por um período se popularizou muito entre os apoiadores deste político¹⁵. Este parece ter sido um dos assuntos comentados neste tópico, como exemplificado no tweet abaixo:

"eu assim reaprendendo a usar às cores do brasil depois de 4 anos de governo do bolsonaro #FIFAWorldCup"

Tabela 20 – Análise de tópicos LDA para a comunidade 2 do dia da estreia da seleção brasileira

Tópico 1	brasil	richarlison	sérvia	fifaworldcup	gols
Tópico 2	neymar	gol	chorando	brasil	richarlison
Tópico 3	hexa	brasil	ganhar	camisa	comeback
Tópico 4	disse	brasil	bolsonaro	falta	senhor
Tópico 5	qatar2022	copa	brasileira	tite	mundo
Tópico 6	rumoaohexa	fifaworldcup	jr	vini	futebol

¹⁴ <https://www.cnnbrasil.com.br/esportes/com-dois-gols-de-richarlison-brasil-vence-a-servia-na-es>

¹⁵ https://www.em.com.br/app/noticia/politica/2022/11/24/interna_politica,1425104/copa-ou-politica-representantes-da-esquerda-vaio-usar-a-camisa-do-brasil.shtml

Na tabela 21 está a análise da comunidade 3, nela foram identificados 2 tópicos. Estes dois tópicos englobam tweets que abordam tweets relacionados a torcida pela seleção brasileira e também ao jogador Neymar. A palavra *liberdade*, que aparece no tópico 2, foi utilizada em um tweet relacionado ao Neymar que recebeu muitos retweets. Este pode ser visto no a seguir:

"Quem é Neymar? Para o cego, é a luz. Para o faminto, é o pão. Para o sedento, é a fonte de água. Para o morto, é a vida. Para o enfermo, é a cura. Para o prisioneiro, é a liberdade. Para o solitário, é o companheiro. Para o viajante, é o caminho. Para mim, é tudo."

Tabela 21 – Análise de tópicos LDA para a comunidade 3 do dia da estreia da seleção brasileira

Tópico 1	brasil	neymar	catar	gol	cultura
Tópico 2	liberdade	povo	lembrar	copadomundo2022	falta

Na Tabela 22 pode ser vista a análise para a comunidade 4. Nesta comunidade foram identificados 4 tópicos. O tópico 1 comenta a respeito do desempenho dos jogadores Neymar e Richarlison nesse primeiro jogo. No tópico 2 é abordado o tema da lesão do Neymar. O terceiro tópico possui relação com tweets que referenciam a torcida pela seleção brasileira. No tópico 4 foi abordado o desempenho do jogador Richarlison, por ter feito dois gols nesse jogo de estreia conforme já foi comentado anteriormente.

Tabela 22 – Análise de tópicos LDA para a comunidade 4 do dia da estreia da seleção brasileira

Tópico 1	neymar	qatar2022	richarlison	fifaworldcup	brasil
Tópico 2	neymar	brasil	chorando	senhor	gol
Tópico 3	tite	brasil	sérvia	hexa	vence
Tópico 4	primeiro	sentiu	richarlison	café	irei

A Tabela 23 ilustra a análise da comunidade 5. Foram identificados 10 tópicos, onde a maioria aborda a lesão do jogador Neymar e a torcida pela seleção brasileira. No tópico 9 aparece *bolsonaro* que está relacionado ao seguinte tweet que foi muito compartilhado:

"e o neymar homenageou o bolsonaro mesmo né fez nada e terminou chorando"

Esse tweet se refere a uma fala de Neymar feita um pouco antes da Copa do Mundo. Neymar disse que comemoraria o primeiro gol na competição referenciando Jair Bolsonaro, presidente em exercício naquele período¹⁶.

¹⁶ <https://oglobo.globo.com/esportes/noticia/2022/10/neymar-diz-que-vai-comemorar-primeiro-gol-na-gh.html>

Tabela 23 – Análise de tópicos LDA para a comunidade 5 do dia da estreia da seleção brasileira

Tópico 1	neymar	brasil	chorando	ganhando	deste
Tópico 2	neymar	nada	chorando	homenageou	terminou
Tópico 3	neymar	disse	gol	tite	primeiro
Tópico 4	pedi	hexa	brasil	jeito	sérvia
Tópico 5	hexa	brasil	rumo	repasse	copa
Tópico 6	senhor	richarlison	saiu	22	lesão
Tópico 7	neymar	jogar	jogador	chora	qatar2022
Tópico 8	campo	sempre	falta	lesionado	silva
Tópico 9	bolsonaro	café	favor	deus	fifaworldcup
Tópico 10	neymar	entender	irei	desculpe	bola

5.4 DIA DA ELIMINAÇÃO DA SELEÇÃO BRASILEIRA NA COMPETIÇÃO

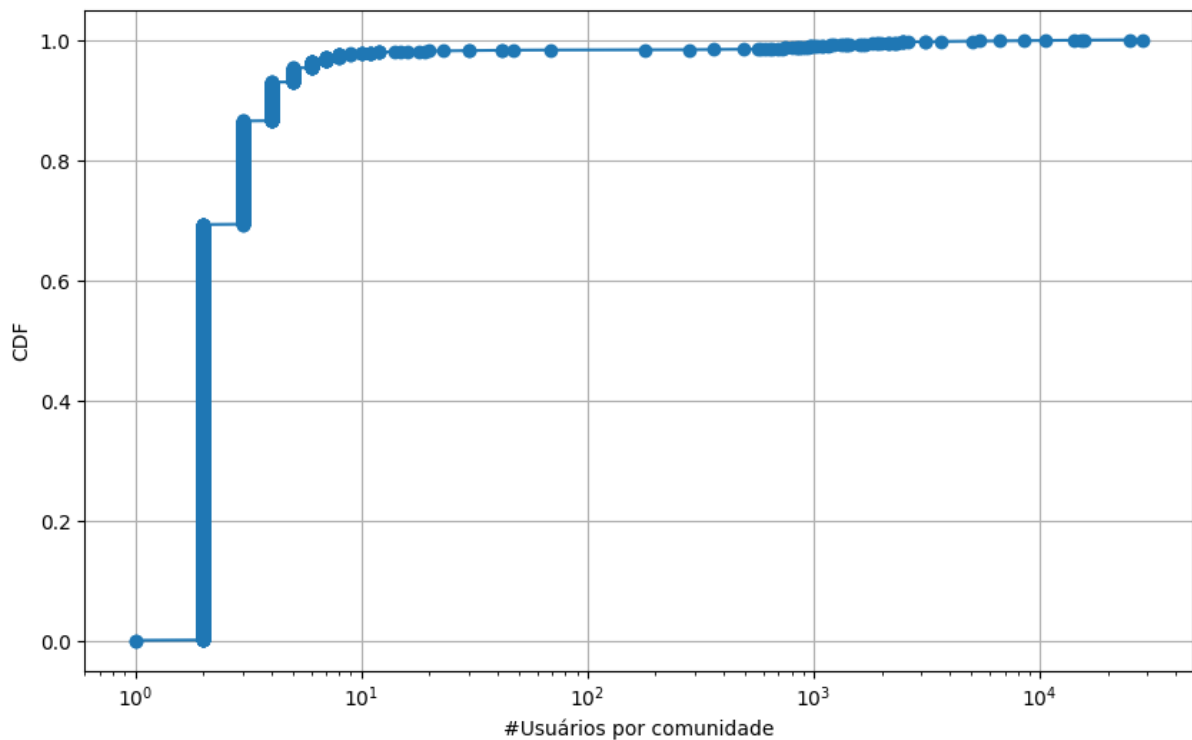
A rede gerada para o dia da eliminação da seleção brasileira abrange 226.491 internautas e revelou 4.011 comunidades. A Tabela 24 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades.

Tabela 24 – Informações gerais sobre os grafos do dia da eliminação da seleção brasileira

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	226.491	14.123.676	124,72	-
Comunidade 1	28.532	643.154	45,08	11
Comunidade 2	24.877	466.058	37,47	10
Comunidade 3	15.769	344.002	43,63	12
Comunidade 4	15.313	271.244	35,43	13
Comunidade 5	14.135	243.236	34,42	11

As comunidades identificadas exibiram uma modularidade de 0,82. A Figura 17 apresenta a distribuição do número de indivíduos por comunidade neste dia, essa distribuição está sendo comparada com os outros dias na seção 5.6 Análise Comparativa.

Figura 17 – Distribuição de número de indivíduos por comunidade nos dia da eliminação da seleção brasileira



A Tabela 25 fornece o total de internautas e a quantidade de retweets em cada comunidade, além dos valores totais para este dia.

Tabela 25 – Número de internautas e retweets por comunidade no dia da eliminação da seleção brasileira

Comunidade	Número de internautas	Número de retweets
1	28.532	58.925
2	24.877	60.649
3	15.769	32.707
4	15.313	32.137
5	14.135	30.584
Todas	226.491	501.062

Perfil dos Indivíduos

Na Tabela 26 está disponível o número de tweets e seguidores de cada um dos 5 indivíduos mais influentes de cada uma das comunidades deste dia, além de mostrar se é ou não verificado. A Tabela 27 possui o mesmo formato, porém mostra os 5 indivíduos mais influentes de cada uma das comunidades.

Tabela 26 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da eliminação da seleção brasileira

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	491	19210	Não	4794	2424889	Sim	161	1514	Não	11	13643	Não	189	43157	Não
2	226	5306	Não	873	11490223	Não	24	33173	Não	6	225510	Não	27	55	Não
3	51	16483	Não	346	116098	Não	22	1958	Não	5	239968	Sim	16	561	Não
4	47	5561	Não	238	156943	Não	11	200927	Sim	5	1270	Não	15	20135	Não
5	25	37845	Não	145	146527	Não	3	3418	Não	4	33355	Não	11	2427	Não

Tabela 27 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da eliminação da seleção brasileira

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	64	9	Não	48	832	Não	83	82	Não	63	8	Não	73	1145	Não
2	54	73	Não	46	814	Não	41	2610	Não	51	52	Não	59	1366	Não
3	48	434	Não	40	189	Não	40	5	Não	33	1593	Não	47	1889	Não
4	45	2907	Não	38	147	Não	30	17	Não	33	598	Não	43	712	Não
5	41	1415	Não	37	391	Não	29	4515	Não	31	26	Não	38	10458	Não

Nuvens dos 50 Pares de Palavras mais Frequentes

A Figura 18 apresenta as nuvens dos 50 pares de palavras. As comunidades 1 e 4 destacam pares de palavras como *obrigado neymar*, *vini jr* e *última copa*. Essas se referem a dúvida que surgiu a respeito desta Copa do Mundo ser a última do Neymar¹⁷ e a boa atuação do jogador Vinicius Jr, que lamentou muito com essa derrota¹⁸. As comunidades 2, 3 e 5 falam a respeito do técnico da seleção brasileira nesta copa. Ele foi muito criticado após a eliminação do Brasil, principalmente por ter saído de campo durante as disputas de pênaltis e não ter feito alterações durante as cobranças¹⁹. Os pares de palavras que se referem a isso são *líder covarde*, *parabéns tite*, *impossível vencer*, *jogo deixando* e *brasil eliminado*.

¹⁷ <https://oglobo.globo.com/esportes/catar-2022/noticia/2022/12/ultima-copa-do-mundo-de-neymar-saiba-o-que-o-craque-ja-disse-sobre-o-mundial-de-2026.ghtml>

¹⁸ <https://ge.globo.com/futebol/selecao-brasileira/noticia/2022/12/10/vinicius-junior-sobre-eliminacao-do-brasil-na-copa-o-pior-dia-da-minha-vida.ghtml>

¹⁹ <https://ge.globo.com/futebol/selecao-brasileira/noticia/2022/12/09/tite-fala-apos-eliminacao-do-brasil-neymar-cobraria-o-quinto-e-decisivo-penalti.ghtml>

Figura 18 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da eliminação da seleção brasileira

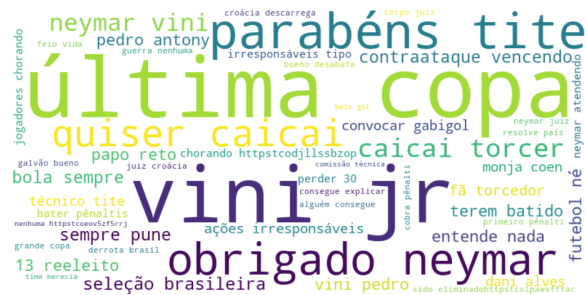


(a) Comunidade 1

(b) Comunidade 2



(c) Comunidade 3



(d) Comunidade 4



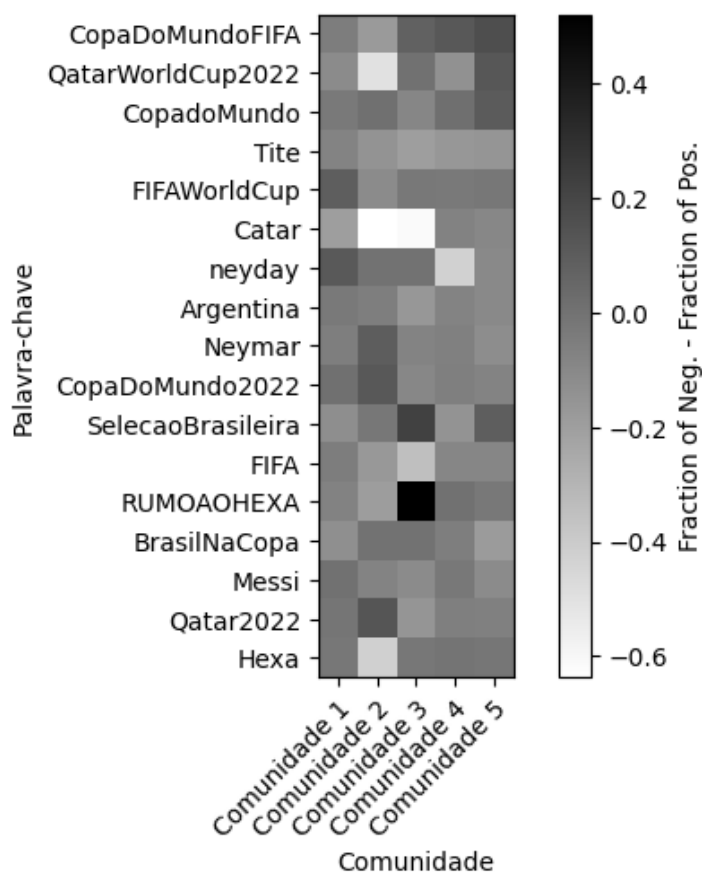
(e) Comunidade 5

Análise de Sentimentos

A análise de sentimento está na Figura 19. De modo geral, os retweets desse dia possuem um sentimento mais neutro para boa parte das palavras chave nas comunidades. Por mais que tenha sido o dia em que a seleção brasileira foi desclassificada da disputa, as palavras chave *FIFAWorldCup* e *neyday* apareceram com sentimentos um pouco positivos na comunidade 1. A comunidade 2 apareceu com sentimentos muito negativos relacionados às palavras-chave *QatarWorldCup2022*, *Catar* e *Hexa*, o que pode demonstrar a insatisfação dos brasileiros com a seleção. Ainda na comunidade 2, as palavras *Neymar* e *CopaDoMundo2022* aparecem com sentimento levemente positivo, o que pode representar uma boa atuação do jogador mesmo na derrota enfrentada. A comunidade 3 está

demonstrando sentimentos muito positivos quanto às palavras *RUMOAOHEXA* e *SelecaoBrasileira*, que pode ser devido aos retweets relacionados a torcida pelo Brasil na partida em que foi eliminado. Nesta comunidade também existe um sentimento muito negativo para a palavra *Catar*. Além disso, a comunidade 4 se destaca com apenas um sentimento mais negativo para a palavra *neyday* e a comunidade 5 com sentimentos positivos para as palavras *CopaDoMundoFIFA*, *QatarWorldCup2022* e *CopadoMundo*.

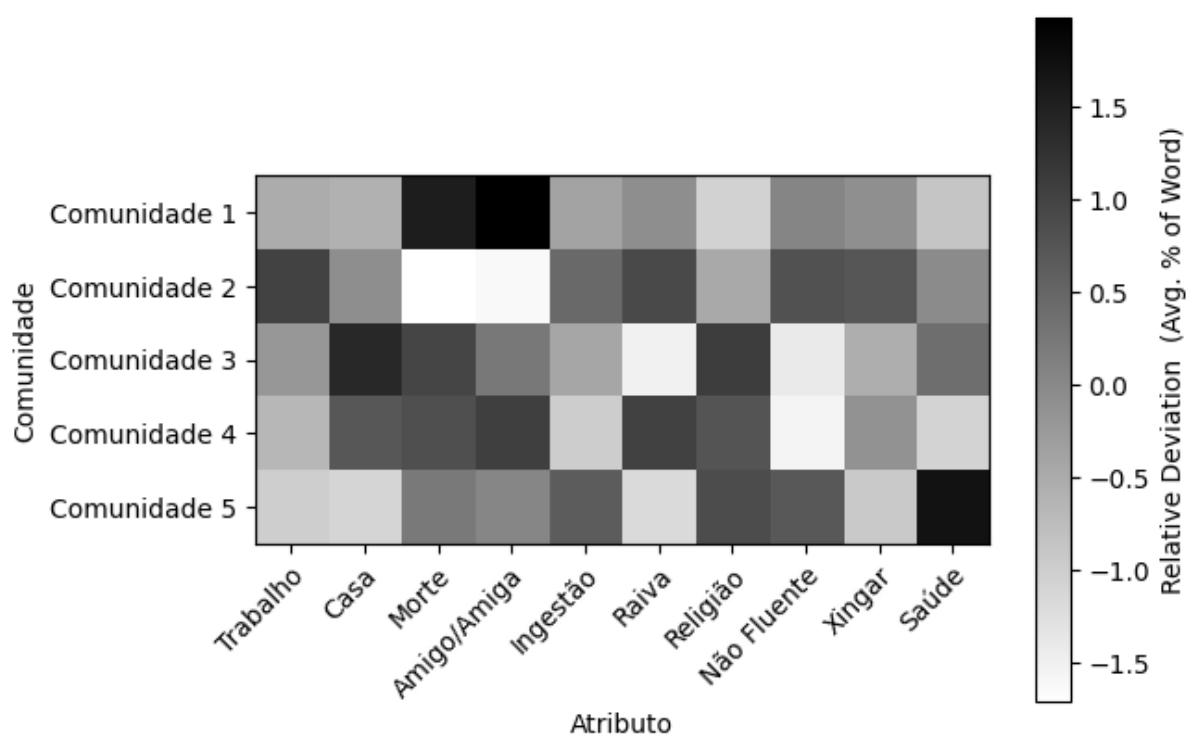
Figura 19 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da eliminação da seleção brasileira



Análise Psicolinguísticas

A Figura 20 apresenta a análise psicolinguística. Na comunidade 1 as palavras estão predominantemente ligadas a temas de amizade e morte. Na comunidade 2 há uma forte associação com palavras relacionadas a trabalho, raiva, xingamentos e não fluência. Nas comunidades 3 e 4 os retweets se concentram em discussões sobre casa, morte e religião, com a comunidade 4 também incluindo amizade e raiva em suas postagens. Por fim, a comunidade 5 se destaca por conter muitas palavras relacionadas à religião, dificuldades de comunicação e, principalmente, questões de saúde.

Figura 20 – Análise de psicolinguística das top 5 maiores comunidades do dia da eliminação da seleção brasileira



Análise de Tópicos LDA

A Tabela 28 possui a análise de tópicos para a comunidade 1 do dia da eliminação da seleção brasileira. Nessa comunidade foram identificados 4 tópicos. Nos tópicos 1, 2 e 3 os assuntos são críticas ao técnico da seleção brasileira no período da Copa do Mundo de 2022, conhecido como Tite. O tweet a seguir exemplifica a insatisfação dos brasileiros com a atuação dele na partida que eliminou a seleção brasileira da competição.

"última coisa que irei twitar sobre essa copa: eu odeio o tite e ele é um sem vergonha"

O tópico 4 faz referência a um tweet que foi muito compartilhado neste dia que elogia o desempenho do jogador Neymar durante a competição, conforme abaixo.

"Podem falar o que quiser, “cai-cai”, torcer por lesão e etc. Só não podem falar que o Neymar não fez história. Eu espero que não seja a última Copa do Mundo dele. Ele merecia muito. Se preparou e deu o seu máximo pra levar o hexa. Não deu Obrigado, Neymar. Te esperamos em 2026."

Tabela 28 – Análise de tópicos LDA para a comunidade 1 do dia da eliminação da seleção brasileira

Tópico 1	neymar	copa	primeiro	mundo	odeio
Tópico 2	tite	brasil	jogadores	campo	seleção
Tópico 3	tite	parabéns	perder	monja	meditação
Tópico 4	futebol	merecia	neymar	erro	cheia

Na Tabela 29 está a análise feita para a comunidade 2, na qual foram identificados 8 tópicos, sendo todos esses baseados em críticas ao técnico da seleção brasileira. Durante a disputa de pênaltis da partida em que o Brasil foi eliminado da competição, o técnico Tite abandonou o campo e foi em direção ao vestiário²⁰. Essa atitude deixou muitos torcedores revoltados, conforme pode ser visto no tweet abaixo:

"Tite foge do campo vergonhosamente. Deveria ter tido aulas com o técnico japonês de como se perde com honra"

No tópico 5 aparece a palavra *lula*, essa palavra faz referência ao candidato que venceu as eleições presidenciais do Brasil em 2022, Luiz Inácio Lula da Silva²¹. No tópico 6 aparece a palavra *petista*, que se refere aos seguidores do partido deste político. Essas palavras aparecem em meio às críticas ao técnico Tite por ser um período muito próximo ao eleitoral e os seguidores do candidato do partido oposto acreditarem que o técnico seria eleitor de Lula²². Abaixo estão disponíveis dois tweets que mostram críticas que fazem o uso dessas palavras:

"Não vai ter mais pão e circo, se o Brasil tivesse levado o hexa iam falar que é por causa do lula, e eu não ia gostar de ver o lula tirando foto com a taça e a seleção! Tchau Brasil, a camisa verde e amarela volta a ser nossa. . . "

"petista Tite correu para o vestiário e deixou o time em campo. Atitude do Tite é de um petista nato..."

²⁰ <https://ge.globo.com/futebol/selecao-brasileira/noticia/2022/12/09/tite-fala-apos-eliminacao-do-brasil-neymar-cobraria-o-quinto-e-decisivo-penalti.ghtml>

²¹ <https://g1.globo.com/politica/eleicoes/2022/noticia/2022/10/30/lula-vence-o-segundo-turno-e-volta-para-o-terceiro-mandato-de-presidente.ghtml>

²² [urlhttps://revistaforum.com.br/esporte/copadomundo/2022/11/24/dia-em-que-tite-se-recusou-cumprimentar-bolsonaro-quando-treinador-parabenizou-lula-127743.html](https://revistaforum.com.br/esporte/copadomundo/2022/11/24/dia-em-que-tite-se-recusou-cumprimentar-bolsonaro-quando-treinador-parabenizou-lula-127743.html)

Tabela 29 – Análise de tópicos LDA para a comunidade 2 do dia da eliminação da seleção brasileira

Tópico 1	comando	digno	técnico	valeu	deixa
Tópico 2	tite	covarde	vestiário	campo	jogadores
Tópico 3	tite	brasil	dupla	nacional	futebol
Tópico 4	líder	tite	excelente	copa	perdido
Tópico 5	tite	lula	seleção	afunda	hino
Tópico 6	falta	petista	mundo	copa	respeito
Tópico 7	parabéns	atitude	único	culpado	tite
Tópico 8	foge	japonês	perde	aulas	vergonhosamente

As análises de tópicos das comunidades 3 e 4 estão disponíveis nas Tabelas 30 e 31 respectivamente. Na comunidade 3 foram identificados 2 tópicos e na comunidade 4 foram 6. Esses tópicos abordam críticas ao técnico da seleção brasileira, conforme já foi comentado nas análises das outras comunidades.

Tabela 30 – Análise de tópicos LDA para a comunidade 3 do dia da eliminação da seleção brasileira

Tópico 1	tite	neymar	bater	copa	primeiro
Tópico 2	tite	brasil	jogadores	futebol	campo

Tabela 31 – Análise de tópicos LDA para a comunidade 4 do dia da eliminação da seleção brasileira

Tópico 1	tite	campo	chorando	neymar	eliminação
Tópico 2	tite	parabéns	pênalti	bater	primeiro
Tópico 3	neymar	nada	papo	13	tirar
Tópico 4	tite	mundo	seleção	copa	vestiário
Tópico 5	tite	jogadores	covarde	neymar	treinador
Tópico 6	odeio	mundo	fraco	futebol	triste

Na análise de tópicos da comunidade 5, disponível na Tabela 32, foram identificados 8 tópicos. Esses tópicos também abordam críticas ao técnico da seleção brasileira, porém no tópico 5 apareceram as palavras *argentina* e *messi*. Isso porque a decisão do técnico Tite de não botar o jogador Neymar para a cobrança do primeiro pênalti das disputas foi comparada a decisão tomada pelo técnico da seleção argentina, que colocou Messi para bater o primeiro pênalti. o tweet abaixo mostra esse cenário:

"messi o primeiro a bater e converter o penalti dando confiança ao time neymar nem teve chance de bater o dele hoje é o dia mais humilhante da vida do suposto treinador tite"

Tabela 32 – Análise de tópicos LDA para a comunidade 5 do dia da eliminação da seleção brasileira

Tópico 1	tite	campo	eliminação	técnico	gabigol
Tópico 2	bater	tite	primeiro	brasil	perder
Tópico 3	tite	neymar	seleção	deixar	treinador
Tópico 4	tite	parabéns	vini	rodrygo	primeiro
Tópico 5	eliminadao	último	argentina	tite	messi
Tópico 6	monja	meditação	errou	culpa	tite
Tópico 7	tite	tirar	mandar	disso	vinicius
Tópico 8	seleção	desculpaa	pedir	brasileira	time

5.5 DIA DA PARTIDA FINAL DA COPA DO MUNDO DE 2022

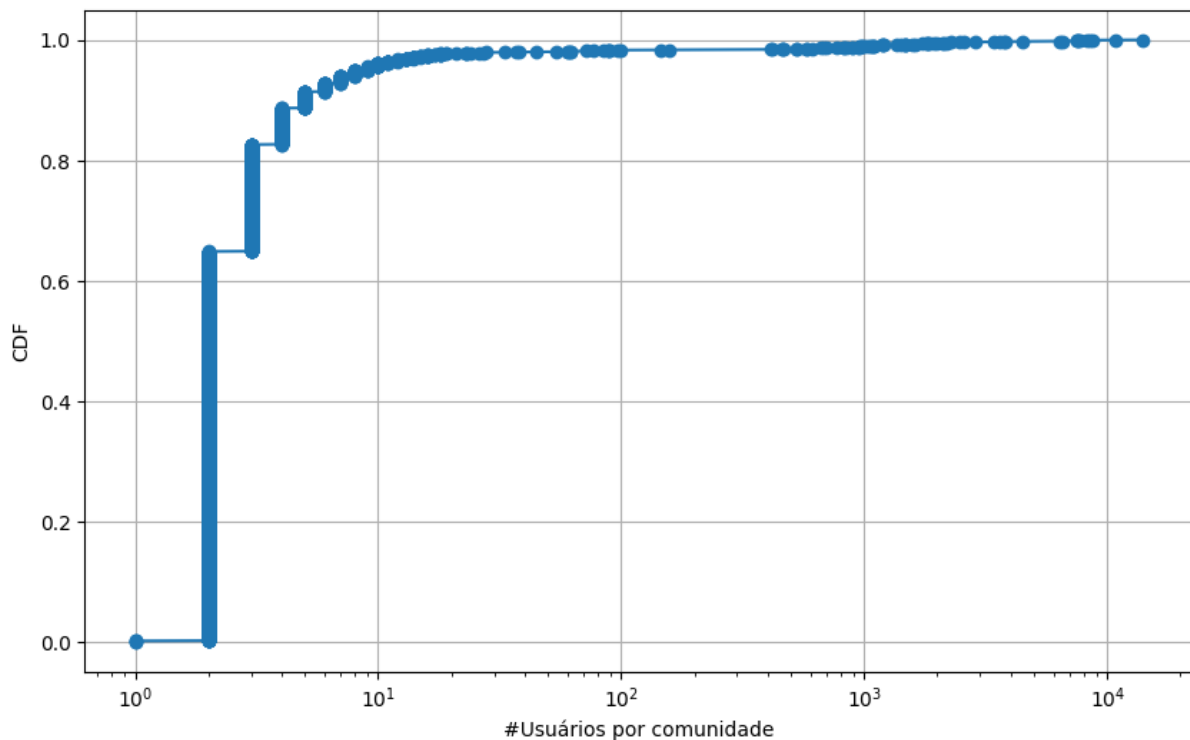
A rede gerada para esse dia da final da competição contou com 193.994 internautas e resultou em 4.298 comunidades. A Tabela 33 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades.

Tabela 33 – Informações gerais sobre os grafos do dia da partida final

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	193.994	11.816.209	121,82	-
Comunidade 1	13.987	442.849	63,32	11
Comunidade 2	10.762	285.500	53,06	11
Comunidade 3	8.679	315.032	72,6	11
Comunidade 4	8.454	180.391	42,68	12
Comunidade 5	8.294	190.430	45,92	11

As comunidades encontradas apresentaram uma modularidade de 0,79. A Figura 21 apresenta a distribuição do número de indivíduos por comunidade neste dia, essa distribuição está sendo comparada com os outros dias na seção 5.6 Análise Comparativa.

Figura 21 – Distribuição de número de indivíduos por comunidade nos dia da partida final



A Tabela 34 fornece o total de internautas e a quantidade de retweets em cada comunidade, além dos valores totais para este dia.

Tabela 34 – Número de internautas e retweets por comunidade no dia da final do evento

Comunidade	Número de internautas	Número de retweets
1	13.987	29.396
2	10.762	26.437
3	8.679	19.818
4	8.454	18.598
5	8.294	19.898
Todas	193.994	461.526

Perfil dos Indivíduos

Na Tabela 35 está disponível o número de tweets e seguidores de cada um dos 5 indivíduos mais influentes de cada uma das comunidades, além de mostrar se é ou não verificado. A Tabela 36 possui o mesmo formato, porém mostra os 5 indivíduos mais influentes de cada uma das comunidades.

Tabela 35 – Informações sobre os 5 indivíduos mais influentes das comunidades do dia da partida final

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	1530	448043	Não	193	315514	Não	87	269228	Não	92	8689	Não	132	33434	Não
2	332	580601	Não	119	22979	Não	74	66623	Não	6	5680	Não	33	51480	Não
3	196	2430534	Sim	74	13426	Não	69	2999	Não	6	38727	Não	29	15585	Não
4	16	169810	Não	17	30000	Não	65	4125	Não	3	2232	Não	22	46587	Não
5	13	35670	Não	11	25763	Não	29	804	Não	3	2092	Não	15	8356	Não

Tabela 36 – Informações sobre os 5 indivíduos mais ativos das comunidades do dia da partida final

	Comunidade 1			Comunidade 2			Comunidade 3			Comunidade 4			Comunidade 5		
	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado	#Tweets	#Seg.	Verificado
1	54	1731	Não	86	720	Não	58	58	Não	29	806	Não	81	108	Não
2	46	3891	Não	59	343	Não	42	6606	Não	28	1966	Não	35	474	Não
3	29	559	Não	42	4394	Não	37	1091	Não	26	99	Não	35	61	Não
4	28	793	Não	31	1089	Não	36	257	Não	23	1065	Não	31	207	Não
5	26	297	Não	30	83	Não	33	59	Não	23	723	Não	29	294	Não

Nuvens dos 50 Pares de Palavras mais Frequentes

Na Figura 22 estão as nuvens dos 50 pares de palavras mais frequentes. Na comunidade 1 é possível ver pares de palavras como *parabéns argentina*, *argentina além*, *companheiros fernando* e *fernando cerimedo*. Isso se refere ao tweet:

"Parabéns Argentina. Além de Messi e seus companheiros Fernando Cerimedo também jogou um bolão. Verdadeiro campeão"

Fernando Cerimedo é um argentino que fez uma live com comentários errôneos sobre as urnas eletrônicas usadas nas eleições brasileiras²³. As comunidades 2, 4 e 5 destacam muitos pares de palavras parabenizando a Argentina e o jogador Messi pela conquista do título do evento, sendo elas *lionel messi*, *melhor jogador*, *messi zerou*, *parabéns argentina* e *primeiro pênalti*. A comunidade 3 faz referência a esse tweet:

"Messi e Mbappé CARREGANDO os seus times de cabo a rabo na Copa e o Neymar lá lacrando pra cima do Casagrande no Instagram e tretando com a Nath Finanças no Twitter"

Nath Finanças é uma influencer que virou assunto no período do evento após questionar o Neymar sobre declaração de imposto de renda, que respondeu insatisfeito²⁴. Neste tweet também fala sobre o Casagrande, que é um comentarista esportivo que fez críticas ao Neymar no período pré copa²⁵.

²³ <https://www.cnnbrasil.com.br/nacional/e-falso-que-existam-urnas-com-15-mil-eleitores/>

²⁴ <https://www.cnnbrasil.com.br/economia/nath-financas-provoca-neymar-em-post-sobre-imposto-de-renda/>

²⁵ <https://www.opovo.com.br/copa-do-mundo-2022/2022/11/08/casagrande-questiona-foco-de-neymar-par-copa/>

Figura 22 – Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da partida final

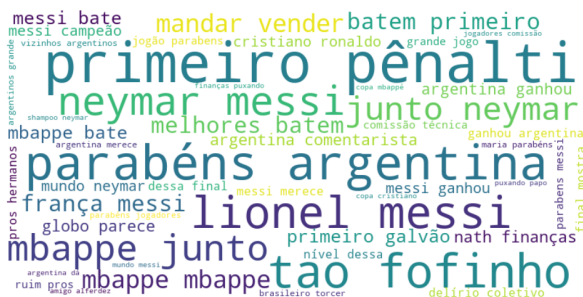


(a) Comunidade 1

(b) Comunidade 2



(c) Comunidade 3



(d) Comunidade 4



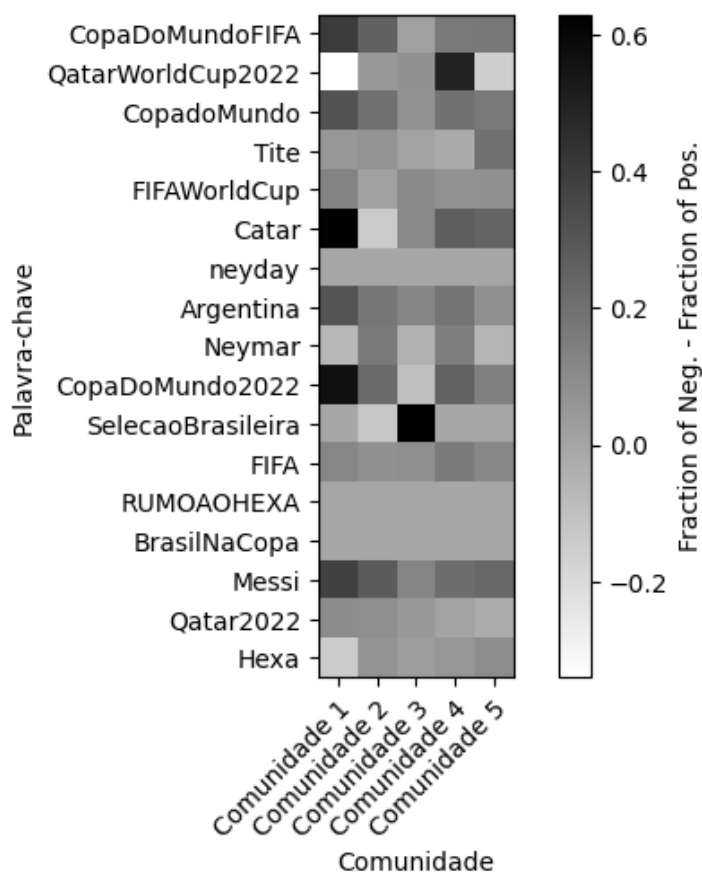
(e) Comunidade 5

Análise de Sentimentos

Na Figura 23 exibe a análise de sentimentos. Nela podem ser vistos sentimentos muito positivos para as palavras-chave *CopaDoMundoFIFA*, *CopadoMundo*, *Catar*, *Argentina* e *Messi* na comunidade 1. Isso pode significar uma satisfação com a partida final, por apresentar Argentina e Messi que representam respectivamente a seleção vencedora do torneio e o jogador camisa 10 dessa seleção, mas também pode representar uma satisfação com o evento como um todo. Ainda nessa comunidade, foi possível observar um sentimento negativo nas palavras *QatarWorldCup2022* e *Hexa*, que pode estar ligado a seleção brasileira não ter chegado a final do evento e consequentemente não ter conquistado o seu sexto título. A comunidade 2 mostra sentimentos negativos para as palavras-chave *Catar*

e *SelecaoBrasileira*. A comunidade 3 apresenta sentimento muito positivo para *SelecaoBrasileira*, mas mostra sentimentos negativos relacionados à *CopaDoMundo2022* e *Neymar*. Já as comunidades 4 e 5 se destacam na palavra *QatarWorldCup2022*, com sentimento positivo na comunidade 4 e negativo na 5.

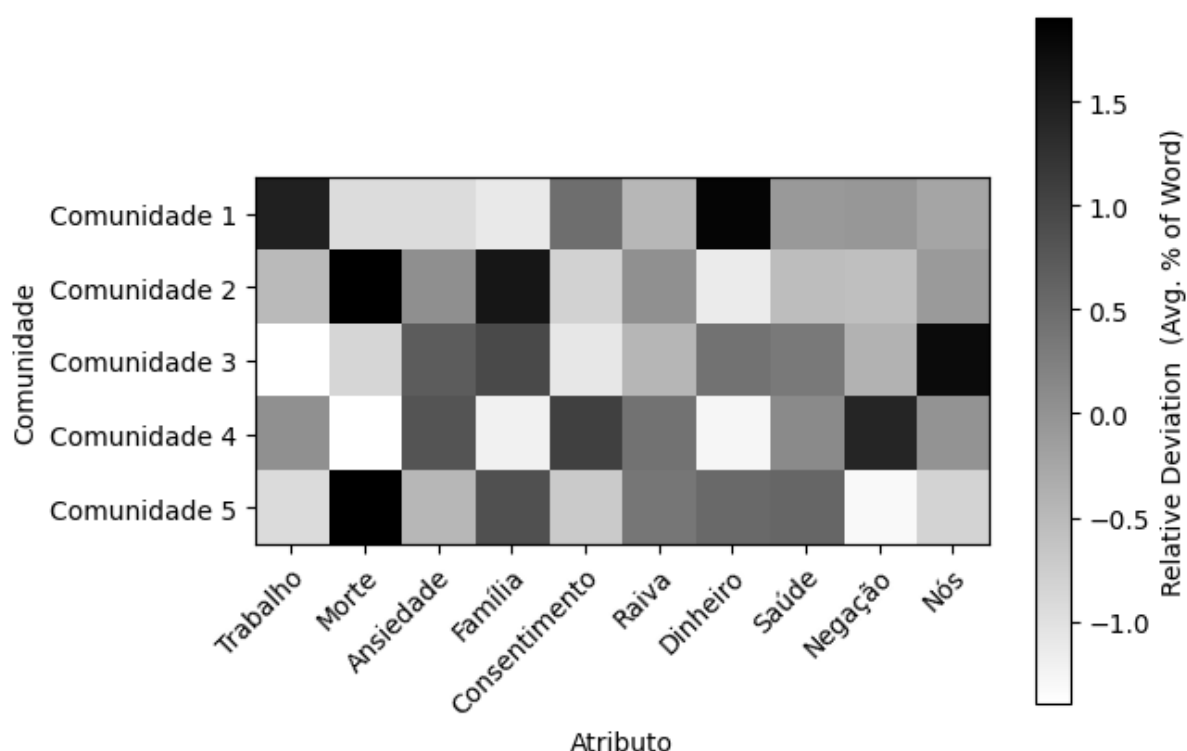
Figura 23 – Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da partida final



Análise Psicolinguísticas

A Figura 24 proporciona uma análise psicolinguística. Na comunidade 1 as palavras predominantes estão relacionadas principalmente a dinheiro e trabalho. Enquanto na comunidade 2 os retweets se concentram mais em questões ligadas à família e à morte. A comunidade 3 reflete a ideia de coletividade, representada pela palavra *nós* na figura. A comunidade 4 se caracteriza por um tom de negação. Por fim, a comunidade 5 mantém uma forte associação com morte.

Figura 24 – Análise de psicolinguística das top 5 maiores comunidades do dia da partida final



Análise de Tópicos LDA

Na tabela 37 está a análise de tópicos para a comunidade 1 do dia da partida final do evento. Nela foram identificados 8 tópicos, sendo que a maioria engloba tweets que comentam sobre o jogo e parabenizam a seleção argentina e o jogador Messi por ter conquistado o torneio. No tópico 2 aparece a palavra *fome*, que é por conta de tweets como este:

"O povo sofrido da Argentina merece essa alegria. A situação econômica é das piores e o povo passa fome. É o que comenta o ex jogador D'Alessandro."

No tópico 6 existem alguns tweets comparando a escolha do técnico da seleção argentina de colocar o melhor jogador do time nesse aspecto para bater o primeiro pênalti das disputas, com a escolha do técnico da seleção brasileira que colocou o melhor jogador para bater o último pênalti. Além disso, existem tweets que se referem a uma fala que o jogador francês Mbappé fez um tempo antes da Copa do Mundo de 2022 falando que o futebol sul-americano não é tão avançado quanto o da Europa²⁶. Segue um tweet que se refere a este tema:

²⁶ <https://www.cnnbrasil.com.br/esportes/em-maio-mbappe-disse-que-futebol-sul-americano-nao-e-tao>

"Quando você descobre que o futebol sul-americano é mais avançado do que pensava... Que pena, Mbappé! #mbappé #frança #copadomundo #argentina #worldcup #final"

No tópico 7, além de falar a respeito do jogo, também houveram tweets que falaram a respeito de manifestações a favor do governo Bolsonaro após a vitória de Lula nas eleições que ocorriam no período do evento²⁷. A palavra *patriotas* neste tópico representa isso. O tweet abaixo mostra como este tema foi citado:

"Comando Militar do Leste no RJ. Enquanto a TV mostrava a final da Copa do Mundo do Catar, milhares de pessoas estavam lutando por liberdade e pela defesa da nossa soberania. Parabéns à todos os Patriotas que saíram às ruas em todo o Brasil!"

Tabela 37 – Análise de tópicos LDA para a comunidade 1 do dia da partida final

Tópico 1	campeão	messi	argentina	cerimedo	fernando
Tópico 2	povo	argentina	merece	fome	jogador
Tópico 3	argentina	técnico	globo	vida	messi
Tópico 4	mundo	parabéns	copa	argentina	chegar
Tópico 5	história	seleção	comando	pessoas	brasil
Tópico 6	tite	aprende	messi	mbappé	argentina
Tópico 7	primeiro	pênalti	messi	defesa	patriotas
Tópico 8	messi	ganhou	garra	série	netflix

A Tabela 38 mostra a análise para a comunidade 2. Nesta comunidade o número de tópicos identificados foi 10. Os tópicos 1, 2, 4, 8, 9 e 10 abordam o desempenho do jogador Messi no torneio e a conquista do título mundial pela Argentina. O tópico 3 fala a respeito do desempenho dos jogadores Neymar, Messi e Mbappé, que atuavam juntos no time Paris Saint-Germain Football Club. No tweet abaixo está um exemplo de tweet que comenta a respeito disso:

"Messi volta ao PSG campeão do Mundo. Mbappé volta ao PSG como artilheiro da Copa e 3 gols na final; também é campeão Mundial. Neymar volta para o PSG MENOR do que chegou ao Mundial e o único que decepcionou entre os craques. Está MUITOS degraus abaixo hoje."

Os tópicos 5, 6 e 7 possuem críticas a escolha do técnico da seleção brasileira por não ter colocado o jogador principal desta função do time para bater primeiro na disputa de pênaltis como foi com a França e a Argentina.

²⁷ <https://www.metropoles.com/colunas/grande-angular/patriotas-convocam-ultimo-ato-apos-45-dias-s>

Tabela 38 – Análise de tópicos LDA para a comunidade 2 do dia da partida final

Tópico 1	melhor	copa	messi	jogador	mundo
Tópico 2	messi	final	mundo	melhor	troféu
Tópico 3	neymar	messi	mbappé	final	copa
Tópico 4	messi	copa	mundo	futebol	lionel
Tópico 5	tite	messi	aprende	título	argentina
Tópico 6	tite	messi	lionel	campeão	seleção
Tópico 7	neymar	aprendeu	messi	copa	jogo
Tópico 8	contra	gol	arg	gols	austrália
Tópico 9	scaloni	primeiro	time	técnico	acordo
Tópico 10	bater	messi	sonho	abriu	partida

Na Tabela 39 disponível a análise da comunidade 3, com 10 tópicos identificados. O tópico 1 fala sobre a situação polêmica já comentada anteriormente a respeito da escolha do técnico da seleção brasileira na escolha da ordem dos jogadores na disputa de pênaltis. O tópico 2 comenta a respeito do desempenho dos jogadores Neymar, Mbappé e Messi no torneio. A discussão do tópico 3 gira em torno da fala do comentarista brasileiro Casagrande, que está disponível no tweet a seguir:

"Casagrande sobre o Brasil na Copa do Mundo: 'Essa final mostraram o quanto o futebol brasileiro está atrasado, fora do cenário mundial. Os dois primeiros batedores foram Mbappé e Messi, e nossos treinadores continuam falando que o melhor tem que ficar por último.'"

O tópico 4 se refere ao seguinte tweet que foi muito retweetado:

"Messi e Mbappé CARREGANDO os seus times de cabo a rabo na Copa e o Neymar lá lacrando pra cima do Casagrande no Instagram e tretando com a Nath Finanças no Twitter"

Os outros tópicos abordam comentários a respeito das conquistas do jogador Messi e também da seleção argentina. No tweet a seguir é possível ver uma fala do jogador Messi que foi muito compartilhada nesta comunidade:

"'É simplesmente inacreditável. Eu sabia que Deus ia me dar a taça, tinha certeza – foi uma grande alegria para nós. Eu tinha esse grande sonho há muito tempo, queria encerrar minha carreira com a Copa do Mundo. Não posso pedir mais do que isso' - Messi"

Tabela 39 – Análise de tópicos LDA para a comunidade 3 do dia da partida final

Tópico 1	tite	copa	mundo	neymar	mbappé
Tópico 2	neymar	messi	instagram	mbappé	qatar2022
Tópico 3	primeiro	casagrande	messi	argentina	frança
Tópico 4	copa	finanças	nath	twitter	neymar
Tópico 5	gaveta	hexa	copa	ouro	taça
Tópico 6	time	copa	live	messi	35
Tópico 7	psg	campeões	mbappé	jr	messi
Tópico 8	chegando	pegar	argentina	copa	futebol
Tópico 9	times	cabo	mbappé	copa	lacrando
Tópico 10	dessa	simplesmente	precisava	sempre	odeiam

As Tabelas 40 e 41 ilustram as análises das comunidades 4 e 5, respectivamente. A comunidade 4 possui 8 tópicos, enquanto a 5 possui 10. Esses tópicos abordam temas já comentados nas outras comunidades, como a conquista do título mundial pela seleção argentina, desempenho dos jogadores Neymar, Mbappé e Messi e polêmica da escolha do da ordem de jogadores pelo técnico da seleção brasileira nos pênaltis. No tópico 8 da comunidade 4 aparece a palavra *bolsonaro*, nome do presidente em exercício durante o torneio, que perdeu as eleições presidenciais do Brasil pouco antes da Copa do Mundo. Um dos tweets que cita ele é o seguinte:

"Bolsonaro sonhava terminar seu ano reeleito recebendo a taça de hexacampeão do Brasil das mãos de Neymar. Acabou com Lula eleito pela terceira vez e a Argentina de Messi, abençoados por Maradona, ser tri-campeões mundiais."

Tabela 40 – Análise de tópicos LDA para a comunidade 4 do dia da partida final

Tópico 1	neymar	mbappé	messi	copa	ganhou
Tópico 2	copa	mundo	messi	lionel	ronaldo
Tópico 3	copa	mundo	messi	final	live
Tópico 4	primeiro	aprende	bate	mbappé	pênalti
Tópico 5	messi	bolado	vender	junto	finanças
Tópico 6	tite	bater	primeiro	mbappé	messi
Tópico 7	messi	argentina	final	lindo	merece
Tópico 8	sempre	bolsonaro	seleção	instagram	parabéns

Tabela 41 – Análise de tópicos LDA para a comunidade 5 do dia da partida final

Tópico 1	messi	copa	mundo	lionel	futebol
Tópico 2	hexa	diferença	tática	scaloni	tite
Tópico 3	copa	andré	mundo	final	tão
Tópico 4	messi	jogo	copa	melhor	parabéns
Tópico 5	tite	ronaldo	cristiano	copa	dessa
Tópico 6	time	copa	messi	argentina	bola
Tópico 7	neymar	brasil	contra	argentina	decepção
Tópico 8	messi	the	fifa	conseguiu	esquecer
Tópico 9	mbappé	primeiro	pênalti	messi	acordo
Tópico 10	quanto	técnico	acabou	2026	batendo

5.6 ANÁLISE COMPARATIVA

Nessa seção foram feitas algumas análises comparando os dias selecionados com alguns dos conteúdos gerados para cada um deles.

Foi possível observar a relação entre o número de internautas e a quantidade de comunidades detectadas, pois isso pode indicar a diversidade de tópicos ou discussões paralelas ocorrendo dentro do contexto principal. Essa comparação foi notável ao analisar o dia da estreia da seleção brasileira em relação ao dia da eliminação. No dia da estreia, haviam 266.622 internautas e 3.527 comunidades detectadas. Em contraste, no dia da eliminação, haviam 226.491 internautas e 4.011 comunidades detectadas. Isso significa que no dia da eliminação foram identificadas mais comunidades, mesmo com um número menor de internautas. Essa tendência também se repete ao comparar o dia da final do evento com esses dias, visto que haviam menos indivíduos, mas um maior número de comunidades detectadas.

Além disso, nas distribuições de número de indivíduos por comunidade disponíveis nas Figuras 9, 13, 17 e 21 foi notável que a maior parte deles está agrupada em um número reduzido de comunidades. Essa observação sugere que entre esses indivíduos há várias discussões paralelas aos principais tópicos em destaque. Nas Tabelas 7, 16, 25 e 34 pode-se notar que a classificação das comunidades com base no número de indivíduos é muito semelhante à classificação com base no número de retweets. A análise dos indivíduos mais influentes e ativos tem como propósito descrever um pouco sobre a distribuição de tweets e retweets dentre as pessoas que estão participando ativamente das discussões sobre a Copa do Mundo de 2022 no Twitter em cada comunidade dos dias selecionados.

Nas Tabelas 17, 17, 26 e 35 estão disponíveis os números de tweets e seguidores de cada um dos 5 indivíduos mais influentes de cada uma das comunidades dos dias selecionados, além de mostrar se é ou não verificado. Nessas tabelas foi possível observar que nas maiores comunidades os indivíduos mais influentes fazem mais tweets do que nas menores. Esses também tendem a ter um número elevado de seguidores, porém existem alguns que possuem um número mais baixo. Dentre os 100 indivíduos influentes

mostrados nestas tabelas nenhum se repetiu entre os dias e apenas 9 são verificados. Dos dias selecionados para a análise, os dias de jogos da seleção brasileira foram os que mais apareceram indivíduos verificados entre os mais influentes, sendo 4 no dia da estreia e 3 no dia da eliminação. É importante ressaltar que essa rede foi construída baseada em retweets, então pode ser que essas pessoas tenham feito os tweets em outros dias, mas estes foram retweetados nos dias analisados.

As Tabelas 9, 18, 27 e 36 mostram os números de retweets e seguidores de cada um dos 5 indivíduos mais ativos de cada uma das comunidades dos dias selecionados, além de mostrar se ele é ou não verificado. Com essas tabelas, foi possível perceber que, diferente dos indivíduos mais influentes, a quantidade de retweets entre os indivíduos mais ativos não muda tanto entre as comunidades como muda entre os influentes. Além disso, nota-se que o número de seguidores entre esses indivíduos mais ativos tende a ser relativamente baixo, especialmente quando comparado aos indivíduos mais influentes, e é importante destacar que nenhum dos cinco indivíduos mais ativos em cada comunidade apareceu em mais de um dia ou possui uma conta verificada.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho com o objetivo de realizar a identificação e análise de redes de opinião a partir de algoritmos de detecção de comunidades, foi realizada uma etapa de caracterização do conjunto de dados para que fossem aplicados esses algoritmos de maneira correta e mais eficiente. Desse modo, foram explorados modelos de análises, como a de perfil dos indivíduos, dos emojis, de sentimentos e psicolinguística. Para este processo foram aplicadas ferramentas como nuvem de palavras, LeIA, LIWC, teste de Kruskal-Wallis e coeficiente de Gini. Dentre as principais dificuldades nesta etapa, destacam-se as análises de sentimento e psicolinguística, uma vez que, comparado com o restante da caracterização de dados, teve uma curva de aprendizado muito maior tanto na parte prática, com as ferramentas utilizadas, como na teórica.

Após a etapa inicial de caracterização do conjunto total dos dados, foi realizada a análise e detecção de comunidades a partir de redes de internautas com retweets comuns geradas. Foram encontradas limitações computacionais que dificultaram este processo. Por conta disso, a abordagem utilizada para contornar essa situação foi a escolha de dias marcantes durante este período para que fosse possível desenvolver o trabalho. Os dias escolhidos foram: o dia da cerimônia de abertura, da estreia da seleção brasileira na competição, da eliminação da seleção brasileira e da partida final da competição.

Com as comunidades já detectadas pelo algoritmo de detecção de comunidades de Louvain, foram escolhidas as cinco maiores comunidades em número de contas para fazer a análise. Dessa forma, na análise dessas comunidades algumas ferramentas já utilizadas na etapa de caracterização geral dos dados foram utilizadas, como LeIA, LIWC, teste de Kruskal-Wallis e coeficiente de Gini. Além disso, foram utilizadas nuvens de pares de palavras, algoritmo LDA e redes de interações de indivíduos para analisar os mais ativos e mais influentes, com essas interações sendo retweets.

Este trabalho mostrou detalhadamente como fazer algumas análises para fazer uma caracterização de dados. Cada uma das análises aplicadas a cada uma das comunidades gerou uma perspectiva diferente a respeito do que estava sendo discutido. Nas nuvens de pares de palavras foi possível observar mais facilmente os assuntos debatidos nos dias selecionados. Já nas análises de sentimento, é possível ter uma perspectiva maior a respeito dos sentimentos aplicados nas palavras escolhidas para as postagens. Na análise psicolinguística também é possível ter uma noção dos sentimentos nas palavras, mas também podemos observar as palavras categorizadas em diferentes atributos relacionados ao estilo linguístico, conceitos afetivos e cognitivos. Com a análise dos indivíduos mais influentes e mais ativos podem ser observadas informações sobre como alguns influenciaram na formação da rede, trazendo a ideia de formadores de opinião.

A análise realizada neste trabalho permite discernir como os eventos externos exercem influência significativa nas interações online, criando extensas redes de indivíduos

alinhados em suas opiniões. Tal compreensão reforça a constatação de que a internet não apenas reflete, mas também impulsiona a propagação de informações. Essa constatação destaca seu papel não apenas como um espelho dos acontecimentos, mas como um substituto vital para os meios tradicionais de comunicação, como jornais e canais de notícias especializados. Assim, é possível compreender sobre a interconexão entre eventos globais e o ambiente virtual, evidenciando a relevância deste estudo para a compreensão do impacto das dinâmicas online na formação de opinião.

No que se refere a trabalhos futuros, são propostas as seguintes ideias:

- Geração da rede do conjunto de dados completo;
- Extração de backbone da rede com todo o conjunto de dados;
- Detecção de comunidades do conjunto de dados completo;
- Comparação de comunidades detectadas com diferentes algoritmos;
- Análise das comunidades detectadas.

REFERÊNCIAS

- ABBADE, Eduardo Botti; DELLA FLORA, Andiará; NORO, Greice de Bem. Interpersonal Influence in Virtual Social Networks and Consumer Decisions. **Revista de Administração da UFSM**, v. 7, n. 2, jun. 2014. DOI: 10.5902/198346594976. Disponível em: <https://periodicos.ufsm.br/reaufsm/article/view/4976>.
- AIRES, Victoria; NAKAMURA, Fabiola. Detecção de Comunidades em Redes Sociais: Relacionando o Método Louvain a Medidas de Centralidade. *In: ANAIS do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC*. São Paulo: SBC, 2017. Disponível em: <https://sol.sbc.org.br/index.php/ctic/article/view/3236>.
- ALMEIDA, Rafael J. A. **LeIA - Léxico para Inferência Adaptada**. [S.l.]: GitHub, 2018. <https://github.com/rafjaa/LeIA>.
- ARAÚJO, Marcelo *et al.* Identificação e Caracterização de Campanhas de Propagandas Eleitorais Antecipadas Brasileiras no Twitter. *In: ANAIS do XII Brazilian Workshop on Social Network Analysis and Mining*. João Pessoa/PB: SBC, 2023. P. 67–78. DOI: 10.5753/brasnam.2023.229879. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/24788>.
- BALAGE FILHO, Pedro P; PARDO, Thiago Alexandre Salgueiro; ALUISIO, Sandra Maria. An evaluation of the Brazilian Portuguese LIWC Dictionary for sentiment analysis. *In: BRAZILIAN Symposium in Information and Human Language Technology - STIL*. [S.l.]: SBC, 2013.
- BELEGANTE, Thaís Caroline; MENEZES, Leonardo Pereira. A influência dos formadores de opinião nas redes sociais. **Anais do 11º ENCITEC 2015**, 2015. Acesso em: 12/06/2023. Disponível em: https://www.fasul.edu.br/projetos/app/webroot/files/controle_eventos/ce_producao/20151027-160644_arquivo.pdf.
- BLEI, David; NG, Andrew; JORDAN, Michael. Latent Dirichlet Allocation. *In: _____*. **Advances in Neural Information Processing Systems**. [S.l.]: MIT Press, 2001. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf.
- CHEN, Emily; LERMAN, Kristina; FERRARA, Emilio. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. **JMIR Public Health Surveill**, v. 6, n. 2, e19273, mai. 2020. ISSN 2369-2960. DOI: 10.2196/19273.
- COELHO, Maurício Archanjo Nunes *et al.* Estratégia Online para Predição Estruturada em Redes Complexas. *In: _____*. **Anais do 11 Congresso Brasileiro de Inteligência Computacional**. Porto de Galinhas, PE: SBIC, 2013. P. 1–6.

- FILHO, José Adail Carvalho; SILVA, Ticiania Linhares Coelho da. Mineração de textos: análise de sentimentos utilizando Tweets referentes à Copa do Mundo 2014, 2014. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/25162>.
- GASTALDO, Édison. "O país do futebol"mediatizado: mídia e Copa do Mundo no Brasil. **SciELO**, Dez. 2009. DOI: 10.1590/S1517-45222009000200013.
- GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences**, v. 99, n. 12, p. 7821–7826, 2002. DOI: 10.1073/pnas.122653799. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.122653799>.
- JAYAWICKRAMA, Thamindu Dilshan. Community Detection Algorithms. **Towards Data Science**, 2021. Acesso em: 12/06/2023. Disponível em: <https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae>.
- KEMP, Simon. **Digital 2021: global overview report**. Acesso em: 27/11/2023. DATAREPORTAL. 2021. Disponível em: <https://datareportal.com/reports/digital-2021-global-overview-report>.
- KRUSKAL, William H.; WALLIS, W. Allen. Use of Ranks in One-Criterion Variance Analysis. **Journal of the American Statistical Association**, Taylor e Francis, v. 47, n. 260, p. 583–621, 1952. DOI: 10.1080/01621459.1952.10483441.
- LINS, Samuel. Preparando-se para a Copa do Mundo: o que leva os brasileiros a comprar impulsivamente produtos para apoiar o seu país? *In*: FPCEUP. ATAS do X Simpósio Nacional de Investigação em Psicologia. [S.l.: s.n.], 2020. Disponível em: <https://hdl.handle.net/10216/128988>.
- MALAGOLI, Larissa *et al.* Caracterização do debate no Twitter sobre a vacinação contra a COVID-19 no Brasil. *In*: ANAIS do X Brazilian Workshop on Social Network Analysis and Mining. Evento Online: SBC, 2021. P. 55–66. DOI: 10.5753/brasnam.2021.16125. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/16125>.
- PAIVA, Beatriz *et al.* O debate do feminismo no Twitter: Um estudo de caso das eleições brasileiras de 2022. *In*: ANAIS do XII Brazilian Workshop on Social Network Analysis and Mining. João Pessoa/PB: SBC, 2023. P. 103–114. DOI: 10.5753/brasnam.2023.230537. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/24791>.
- PENFOLD, Tom. National identity and sporting mega-events in Brazil. **Sport in Society**, Routledge, v. 22, n. 3, p. 384–398, 2019. DOI: 10.1080/17430437.2018.1490266.

SANTIAGO, Rafael de. **Anotações para a Disciplina de Grafos**. [S.l.: s.n.], 2023. Acesso em: 15/06/2023. Disponível em: www.inf.ufsc.br/~r.santiago/downloads/INE5413.pdf.

TAUSCZIK, Yla R.; PENNEBAKER, James W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. **Journal of Language and Social Psychology**, v. 29, n. 1, p. 24–54, 2010. DOI: 10.1177/0261927X09351676. Disponível em: <https://doi.org/10.1177/0261927X09351676>.

TEIXEIRA, Diogo; AZEVEDO, Isabel. Análise de opiniões expressas nas redes sociais. **Revista Ibérica de Sistemas e Tecnologias de Informação**, n. 8, 2011. ISSN 1646-9895. DOI: 10.4304/risti.8.53-65.

VASCONCELLOS-SILVA, Paulo; ARAÚJO-JORGE, Tânia. Análise de conteúdo por meio de nuvem de palavras de postagens em comunidades virtuais: novas perspectivas e resultados preliminares. *In*: ANAIS do 8º Congresso Ibero-Americano em Investigação Qualitativa. Lisboa, Portugal: CIAIQ, 2019. P. 41–48. Disponível em: <https://proceedings.ciaiq.org/index.php/CIAIQ2019/article/view/2002>.

VOLPATO, Bruno. **Ranking: as redes sociais mais usadas no Brasil e no mundo em 2023, com insights, ferramentas e materiais**. Acesso em: 12/06/2023. Resultados Digitais. 2023. Disponível em: <https://resultadosdigitais.com.br/marketing/redes-sociais-mais-usadas-no-brasil/>.

YITZHAKI, Shlomo. Relative Deprivation and the Gini Coefficient. **The Quarterly Journal of Economics**, v. 93, n. 2, p. 321–324, mai. 1979. ISSN 0033-5533. DOI: 10.2307/1883197.

Apêndices

APÊNDICE A – CÓDIGO

```

import pandas as pd
from gensim.corpora import Dictionary
from gensim.models import LdaModel
from matplotlib import pyplot as plt
import matplotlib.cm as cm
import numpy as np
import collections
import emoji
import os.path
import csv
import io
import pytz
import matplotlib.dates as mdates
from wordcloud import WordCloud
from collections import Counter
import re
import nltk
from nltk.corpus import stopwords
from nltk.util import ngrams
from nltk.tokenize import RegexpTokenizer
from datetime import datetime
from gensim import corpora, models
import pyLDAvis.gensim_models as gensimvis
import pyLDAvis
from unwanted_words import unwanted_words
import gensim
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import strip_punctuation
from gensim.models import CoherenceModel
from liwc import LIWC, attribute_translation
from scipy.stats import kruskal
from scipy.cluster.hierarchy import dendrogram, linkage
from matplotlib.colors import LogNorm
import networkx as nx
from collections import defaultdict
from networkx.algorithms.community import girvan_newman
import tqdm
import community

```

```

csv_limiter = ;
stop_words = stopwords.words('portuguese')
liwc = LIWC('LIWC2007_Portugues_win.dic')
keywords_data = [
    'CopaDoMundoFIFA',
    'QatarWorldCup2022',
    'CopaDoMundo',
    'CopaDoMundoFIFA',
    'Tite',
    'FIFAWorldCup',
    'Catar',
    'neyday',
    'Argentina',
    'Neymar',
    'CopaDoMundo2022',
    'SelecaoBrasileira',
    'FIFA',

```

```

    'RUMOAOHEXA',
    'BrasilNaCopa',
    'Messi',
    'Qatar2022',
    'Hexa'
]

# Conta o numero de tweets/retweets.
def count_tweets(table: pd.DataFrame) -> int:
    return table.shape[0]

# Conta o numero de tweets/retweets por palavra-chave.
def count_tweets_or_retweet_per_keyword(keywords_dir: str, is_retweet: bool) -> dict:
    # Define o tamanho maximo do campo em 100 MB.
    csv.field_size_limit(100000000)
    num_tweets_per_keyword = {}
    for keyword in os.listdir(keywords_dir):
        current_dir = keywords_dir + '/' + keyword
        for file in os.listdir(current_dir):
            if is_retweet:
                if file.endswith('.csv') and 'RT' in file:
                    with open(os.path.join(current_dir, file), 'rb') as file:
                        content = file.read()
                        content = content.replace(b'\x00', b'')
                        csv_file = io.StringIO(content.decode('utf-8'))
                        csv_reader = csv.reader(csv_file, delimiter='csv_limiter')
                        num_linhas = sum(1 for _ in csv_reader)
                        if keyword in num_tweets_per_keyword:
                            num_tweets_per_keyword[keyword] += num_linhas
                        else:
                            num_tweets_per_keyword[keyword] = num_linhas
            else:
                if file.endswith('.csv') and 'RT' not in file:
                    with open(os.path.join(current_dir, file), 'rb') as file:
                        content = file.read()
                        content = content.replace(b'\x00', b'')
                        csv_file = io.StringIO(content.decode('utf-8'))
                        csv_reader = csv.reader(csv_file, delimiter='csv_limiter')
                        num_linhas = sum(1 for _ in csv_reader)
                        if keyword in num_tweets_per_keyword:
                            num_tweets_per_keyword[keyword] += num_linhas
                        else:
                            num_tweets_per_keyword[keyword] = num_linhas
    return num_tweets_per_keyword

# Conta o numero de tweets/retweets por semanas de cada palavra-chave.
def count_tweets_or_retweet_per_keyword_and_weeks(
    keywords_dir: str,
    is_retweet: bool,
    begin: int = 0,
    end: int = 8
) -> dict:
    weeks = {'Semana_'} + str(i + 1): {} for i in range(begin, end + 1)}
    # Define o tamanho maximo do campo em 100 MB.
    csv.field_size_limit(100000000)
    for keyword in os.listdir(keywords_dir):
        current_dir = keywords_dir + '/' + keyword + '/'

```

```

for i in range(begin, end + 1):
    if is_retweet:
        dataset_name = 'dataset_RT_info_' + str(i) + '.csv'
    else:
        dataset_name = 'dataset_info_' + str(i) + '.csv'
    week = 'Semana_' + str(i + 1)
    current_dataset_name = current_dir + dataset_name
    if os.path.exists(current_dataset_name):
        with open(current_dataset_name, 'rb') as file:
            content = file.read()
            content = content.replace(b'\x00', b'')
            csv_file = io.StringIO(content.decode('utf-8'))
            csv_reader = csv.reader(csv_file, delimiter='csv_limiter')
            num_linhas = sum(1 for _ in csv_reader)
            weeks[week][keyword] = num_linhas
    else:
        weeks[week][keyword] = 1

return weeks

# Le o csv convertendo a coluna date em um formato de data e
# eliminando linhas com valores nulos.
def read_csv(path: str) -> pd.DataFrame:
    df = pd.read_csv(
        path,
        on_bad_lines='skip',
        dtype='unicode',
        delimiter='csv_limiter',
        engine="python",
        parse_dates=['date']
    )
    df['date'] = pd.to_datetime(df['date'], errors='coerce')
    df.dropna(subset=['date'], inplace=True)
    return df

# Gera o grafico de volume de tweets e retweets durante o periodo coletado.
def generates_graph_of_the_volume_of_tweets_and_retweets_during_collected_period(
    dataset_tweets: pd.DataFrame,
    dataset_retweets: pd.DataFrame,
    point_dates: list = None,
    title: str = '',
    xlabel: str = '',
    ylabel: str = ''
) -> None:
    counts_RT = dataset_retweets.groupby(
        pd.Grouper(key='date', freq='D')).size()
    counts_T = dataset_tweets.groupby(pd.Grouper(key='date', freq='D')).size()

    # Plotar o grafico de linhas com Matplotlib.
    plt.plot(counts_RT.index, counts_RT.values, label='Retweets')
    plt.plot(counts_T.index, counts_T.values, label='Tweets')

    ymin, ymax = plt.axis()[2:4]
    ylen = ymax - ymin
    space_label_dot = (ylen * 2) / 100

    # Converte as datas em objetos pandas.Timestamp com fuso horario.
    point_dates = [pd.Timestamp(date, tz=pytz.UTC) for date in point_dates]
    point_labels = [str(i + 1) for i in range(0, len(point_dates) + 1)]

```

```

# Adicionar pontos na linha 'Tweets'.
for date in point_dates:
    if date in counts_RT.index:
        plt.scatter(date, counts_RT[date], marker='o', color='blue')
        plt.text(date, counts_RT[date] +
                 space_label_dot, point_labels.pop(0))

# Formatar as datas
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%d/%m/%Y'))

# Adicionar legenda
plt.legend()

# Adicionar titulos e rotulos
plt.title(title)
plt.xlabel(xlabel)
plt.ylabel(ylabel)

plt.xticks(rotation='vertical')

# Mostrar o grafico
plt.show()

# Gera a nuvem de palavras mais populares (top-100) dos tweets e retweets.
def generates_cloud_of_the_100_most_popular_words_from_tweets(
    dataset: pd.DataFrame,
    keywords: list = None
) -> None:
    if keywords:
        # Filtra os textos que contem as palavras-chave.
        filtered_dataset = dataset[dataset['text'].str.contains(
            '|'.join(keywords), case=False)]

        text = ' '.join(filtered_dataset['text'])
    else:
        text = ' '.join(dataset['text'])

    text = re.sub(r'[^a-zA-Z0-9]', '', text)

    # Divide a string em palavras.
    words = nltk.word_tokenize(text)

    # Transforma todas as palavras em minusculas.
    words_lower = [str(word).lower() for word in words]

    # Remove as stopwords.
    words_filtered = [word for word in words_lower if word not in stop_words]

    # Conta a frequencia de cada palavra.
    word_freq = Counter(words_filtered)

    # Seleciona as top 100 palavras mais populares.
    top_words = word_freq.most_common(100)

    lista_str = [tupla[0] for tupla in top_words]
    lista_str = list(set(lista_str))
    lista_str.sort(key=len)
    print(lista_str)

```

```

# Cria a nuvem de palavras.
wordcloud = WordCloud(
    width=800,
    height=400,
    background_color='white'
).generate_from_frequencies(dict(top_words))
plt.figure(figsize=(12, 10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()

def generates_cloud_of_the_50_most_popular_bigrams_from_tweets(
    dataset: pd.DataFrame,
    keywords: list = None
) -> None:
    if keywords:
        # Filtra os textos que contem as palavras-chave.
        filtered_dataset = dataset[dataset['text'].str.contains(
            '|'.join(keywords), case=False)]

        text = ' '.join(filtered_dataset['text'])
    else:
        text = ' '.join(dataset['text'])

    text = re.sub(r'^a-zA-Z0-9', '', text)

    # Divide o texto em bigramas.
    words = nltk.word_tokenize(text)
    bigrams = list(ngrams(words, 2))

    # Transforma todas as palavras em minusculas.
    bigrams_lower = [' '.join(bigram).lower() for bigram in bigrams]

    # Remove os bigramas que contem stopwords.
    bigrams_filtered = [bigram for bigram in bigrams_lower if all(
        word not in stop_words for word in bigram.split())]

    # Conta a frequencia de cada bigrama.
    bigram_freq = Counter(bigrams_filtered)

    # Seleciona os top 50 bigramas mais populares.
    top_bigrams = bigram_freq.most_common(50)

    lista_str = [tupla[0] for tupla in top_bigrams]
    lista_str = list(set(lista_str))
    lista_str.sort(key=len)
    print(lista_str)

    # Cria a nuvem de palavras com bigramas.
    wordcloud = WordCloud(
        width=800,
        height=400,
        background_color='white'
    ).generate_from_frequencies(dict(top_bigrams))

    plt.figure(figsize=(12, 10))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')

```

```

plt.show()

# Pega o sentimento de tweets/retweets por semanas de cada palavra-chave.
def get_sentiment_tweets_or_retweet_per_weeks_and_keywords(
    keywords_dir: str,
    is_retweet: bool,
    begin: int = 0,
    end: int = 8
) -> dict:
    weeks = {'Semana_': str(i + 1): {} for i in range(begin, end + 1)}
    # Define o tamanho maximo do campo em 100 MB.
    csv.field_size_limit(100000000)
    for keyword in os.listdir(keywords_dir):
        current_dir = keywords_dir + '/' + keyword + '/'
        for i in range(begin, end + 1):
            if is_retweet:
                dataset_name = 'dataset_RT_info_' + str(i) + '.csv'
            else:
                dataset_name = 'dataset_info_' + str(i) + '.csv'
            week = 'Semana_' + str(i + 1)
            current_dataset_name = current_dir + dataset_name
            if os.path.exists(current_dataset_name):
                dataset = read_csv(current_dataset_name)
                sentiment_polarity_column = dataset['sentiment_polarity'].astype(
                    float)
                sentiment_polarity = sentiment_polarity_column.mean()
                weeks[week][keyword] = sentiment_polarity
            else:
                weeks[week][keyword] = 0

    return weeks

# Gera grafico heatmap de tweets ou retweets por semanas e palavras-chave.
def generates_heatmap_graph_of_tweets_or_retweets_per_weeks_and_keywords(
    data_tweets_per_week_and_keyword: dict,
    cmap: str = 'RdYlGn',
    title: str = '',
    x_label: str = '',
    y_label: str = '',
    label_color_bar: str = '',
    log_scale: bool = False
) -> None:
    keywords = list(data_tweets_per_week_and_keyword[list(
        data_tweets_per_week_and_keyword.keys())[0]].keys())
    weeks = list(data_tweets_per_week_and_keyword.keys())
    data_tweets_per_keyword = [
        list(week.values()) for week in data_tweets_per_week_and_keyword.values()]
    data_tweets_per_keyword_inverted = [
        list(tupla) for tupla in zip(*data_tweets_per_keyword)]

    # Crie um array de dados.
    dados = np.array(data_tweets_per_keyword_inverted)

    # Crie o grafico de heatmap usando a funcao imshow do Matplotlib.
    fig, ax = plt.subplots()
    if log_scale:
        heatmap = ax.imshow(dados, cmap=cmap, norm=LogNorm())
    else:

```

```

heatmap = ax.imshow(dados, cmap=cmap)

# Adicione a barra de cores.
cax = ax.figure.colorbar(heatmap, ax=ax)

# Defina o texto da legenda.
cax.set_label(label_color_bar)

# Adicione as etiquetas dos eixos.
ax.set_xticks(np.arange(dados.shape[1]))
ax.set_yticks(np.arange(dados.shape[0]))
ax.set_xticklabels(weeks)
ax.set_yticklabels(keywords)

# Rotacione os rotulos do eixo x.
plt.setp(ax.get_xticklabels(), rotation=45,
          ha="right", rotation_mode="anchor")

# Adicione o titulo.
ax.set_title(title)

# Adicione a label do eixo x.
ax.set_xlabel(x_label)

# Adicione a label do eixo y.
ax.set_ylabel(y_label)

# Mostre o grafico.
plt.show()

# Gera o grafico de distribuicoes dos numeros de tweets e
# retweets por tipo de conta de usuario.
def generates_distribution_chart_of_numbers_tweets_and_retweets_by_type_of_user_account(
    dataset: pd.DataFrame,
    title: str = '',
    xlabel: str = '',
    ylabel: str = ''
) -> None:
    # Calculo do numero de tweets por usuario
    count_by_user = dataset.groupby('screen_name').size()

    # Separacao dos usuarios verificados e nao verificados
    verified = count_by_user[count_by_user.index.isin(
        dataset[dataset['verified'] == 'True']['screen_name'])]
    not_verified = count_by_user[count_by_user.index.isin(
        dataset[dataset['verified'] == 'False']['screen_name'])]

    # Calculo da distribuicao acumulada de frequencia (CDF)
    cdf_verified = np.cumsum(
        verified.value_counts(normalize=True).sort_index())

    print('CDF Verificado:')
    for i, cdf in enumerate(cdf_verified.index):
        if i == 80:
            break
        print(f'index: {cdf} - cdf: {cdf_verified[cdf]}')

    cdf_not_verified = np.cumsum(
        not_verified.value_counts(normalize=True).sort_index())

```

```

print( 'CDF_Nao_Verificado:')
for j, cdf in enumerate(cdf_not_verified.index):
    if j == 80:
        break
    print(f'index: {cdf}: {cdf_not_verified[cdf]}')

# Plotagem do grafico
plt.plot(cdf_verified.index, cdf_verified, label='Verificados')
plt.plot(cdf_not_verified.index, cdf_not_verified, label='Nao verificados')
plt.xscale('log')
plt.xlabel(xlabel)
plt.ylabel(ylabel)
plt.title(title)
plt.legend()
plt.show()

# Conta o numero de emojis no texto e retorna um counter com os
# emojis e o numero de vezes que foram utilizados.
def count_emojis(text: str) -> Counter:
    emoji_list = [char for char in text if char in emoji.UNICODE_EMOJI]
    return collections.Counter(emoji_list)

# Gera o grafico dos 10 emojis mais utilizados em tweets ou retweets.
def generates_graph_of_the_10_most_frequent_emojis_in_tweets_or_retweets(
    dataset: pd.DataFrame,
    title: str = '',
    xlabel: str = '',
    ylabel: str = ''
) -> None:
    dataset['emoji_count'] = dataset['text'].apply(count_emojis)

    total_rows = len(dataset)
    all_emojis = collections.Counter()
    for emoji_count in dataset['emoji_count']:
        all_emojis += emoji_count

    top_n = 10
    top_emojis = all_emojis.most_common(top_n)

    emojis = [e for e, _ in top_emojis]
    counts = [count / total_rows * 100 for _, count in top_emojis]

    plt.bar(emojis, counts)
    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.show()

print( "Lista de emojis mais utilizados: ")
for emoji_unicode in emojis:
    emoji_char = emoji.emojize(emoji_unicode)
    print(emoji_char, end=' ')

# Faz analise de topicos LDA.
def lda_topic_analysis(
    dataset: pd.DataFrame,
    num_topics: int = 5,

```



```

        num_iterations: int = 100
    ) -> tuple[LdaModel, list[list[tuple[int, int]]], Dictionary, int]:
        tokenizer = RegexpTokenizer(r'\w+')

    def preprocess(text):
        # Pre-processamento do texto
        tokens = tokenizer.tokenize(text.lower())
        tokens = [token for token in tokens if token not in stop_words]
        return tokens

    dataset['processed_text'] = dataset['text'].apply(preprocess)

    # Criar o dicionário de palavras e o corpus
    dictionary = corpora.Dictionary(dataset['processed_text'])
    corpus = [dictionary.doc2bow(text) for text in dataset['processed_text']]

    lda_model = models.LdaModel(
        corpus=corpus,
        id2word=dictionary,
        num_topics=num_topics,
        iterations=num_iterations
    )
    return lda_model, corpus, dictionary, num_topics

# Gera uma interface interativa com a análise de tópicos LDA.
def generates_interactive_interface_with_LDA_topic_analysis(
    dataset: pd.DataFrame,
    num_topics: int = 5,
    num_iterations: int = 100
) -> None:
    lda_model, corpus, dictionary, _ = lda_topic_analysis(
        dataset=dataset,
        num_topics=num_topics,
        num_iterations=num_iterations
    )
    vis = gensimvis.prepare(lda_model, corpus, dictionary)
    return pyLDAvis.display(vis)

# Pega o número de ocorrência das palavras nos tweets.
def get_occurrence_of_words(dataset: pd.DataFrame, words_list: list) -> dict:
    text_column = dataset['text'].str.lower()
    occurrence = {}
    for word in words_list:
        occurrence[word] = text_column.str.count(word).sum()
    return occurrence

# Gera tabela com o número de ocorrência das palavras nos tweets.
def generates_occurrence_of_words_table(
    dataset: pd.DataFrame,
    words_list: list
) -> pd.DataFrame:
    occurrence = get_occurrence_of_words(dataset, words_list)
    data_table = {'Palavra': occurrence.keys(),
                  'Ocorrencia': occurrence.values()}
    return pd.DataFrame(data_table)

```

```

# Pega os cinco tweets com mais likes que possuem as palavras.
def get_example_of_tweets_with_specific_words(
    dataset: pd.DataFrame,
    words_list: list
) -> pd.DataFrame:
    popular_tweets = {}
    for word in words_list:
        filtered_tweets = dataset[dataset['text'].str.contains(word)]
        ordered_tweets = filtered_tweets.sort_values(
            by='likes', ascending=False)
        top_tweets = ordered_tweets.head(5)
        popular_tweets[word] = top_tweets['text'].to_list()
    return popular_tweets

# Gera tabela com os dois tweets com mais likes que possuem as palavras.
def generates_example_of_tweets_with_specific_words_table(
    dataset: pd.DataFrame,
    words_list: list,
    table_name: str = ''
) -> pd.DataFrame:
    occurrence = get_example_of_tweets_with_specific_words(dataset, words_list)
    rows = []

    for word, tweets in occurrence.items():
        for i, tweet in enumerate(tweets, start=1):
            rows.append({"Palavra": word, "Tweet": tweet})

    table = pd.concat([pd.DataFrame([row]) for row in rows], ignore_index=True)
    table.to_csv('media/tables/' + table_name + '.csv', index=False)
    return table

# Pega as pontuacoes de coerencia pelos numeros de topicos
def get_consistency_by_number_of_topics_lda(
    dataset: pd.DataFrame,
    begin: int = 2,
    end: int = 10,
    step_size: int = 2
) -> None:
    def remove_stopwords(text):
        words_lower = text.lower().split()
        return [word for word in words_lower if word not in stop_words]

    # Pre-processar os dados
    def preprocess(text):
        words_filtered = remove_stopwords(text)
        text = ' '.join(words_filtered)
        text = strip_punctuation(text)
        tokens = simple_preprocess(text, deacc=True)
        return tokens

    dataset['tokens'] = dataset['text'].apply(preprocess)
    # Criar dicionario e corpus
    id2word = Dictionary(dataset['tokens'])
    # Remover palavras de baixa frequencia e muito frequentes
    id2word.filter_extremes(no_below=5, no_above=0.5)
    corpus = [id2word.doc2bow(tokens) for tokens in dataset['tokens']]

    # Executar varios modelos LDA e calcular a coerencia

```

```

coherences = []
for num_topics in range(begin, end+1, step_size):
    lda_model = LdaModel(
        corpus=corpus,
        id2word=id2word,
        num_topics=num_topics,
        random_state=100,
        update_every=1,
        passes=10,
        alpha='auto',
        per_word_topics=True
    )
    coherence_model = CoherenceModel(
        model=lda_model, texts=dataset['tokens'], dictionary=id2word, coherence='c_v'
    )
    coherence_score = coherence_model.get_coherence()
    print(
        f"Pontuacao de coerencia para {num_topics} topicos: {coherence_score}")
    coherences.append(coherence_score)

def generate_liwc_analysis_of_datasets_by_keyword(keywords_dir: str) -> dict:
    liwc_analysis = {}
    for keyword in os.listdir(keywords_dir):
        folder_path = os.path.join(keywords_dir, keyword)
        if os.path.isdir(folder_path):
            csv_files = [arquivo for arquivo in os.listdir(
                folder_path) if arquivo.endswith('.csv')]
            if csv_files:
                datasets = []
                for csv_file in csv_files:
                    csv_file_path = os.path.join(folder_path, csv_file)
                    dataset = read_csv(csv_file_path)
                    datasets.append(dataset)
                dataset_keyword = pd.concat(datasets)
                liwc_dataset_keyword = liwc.process_df(dataset_keyword, 'text')
                liwc_analysis[keyword] = liwc_dataset_keyword
    return liwc_analysis

# Aplica o teste nao parametrico de Kruskal na lista de analises LIWC para pegar
# os atributos que possuem uma diferenca significativa.
def apply_kruskal_test_in_liwc_analysis(liwc_analysis_list: list) -> list:
    liwc_analysis_list = [liwc_analysis.fillna(
        0) for liwc_analysis in liwc_analysis_list]
    selected_attributes = []
    for i in range(len(liwc_analysis_list)):
        for j in range(len(liwc_analysis_list)):
            if i == j:
                continue
            for attribute in liwc_analysis_list[i].columns:
                _, p_value = kruskal(
                    liwc_analysis_list[i][attribute], liwc_analysis_list[j][attribute])
                significance_level = 0.05
                if p_value < significance_level:
                    selected_attributes.append(attribute)
    selected_attributes = list(set(selected_attributes))
    return selected_attributes

```

```

# Gera a analise LIWC para o dataset.
def generate_liwc_analysis_for_dataset(dataset: pd.DataFrame) -> pd.DataFrame:
    liwc = LIWC('LIWC2007_Portugues_win.dic')
    liwc_analysis = liwc.process_df(dataset, 'text')
    liwc_analysis = liwc_analysis.fillna(liwc_analysis.mean())
    return liwc_analysis

# Gera tabela com a media das analises LIWC por palavra-chave.
def generate_table_with_average_analysis_liwc_by_keyword(
    liwc_analysis_per_keyword: dict
) -> pd.DataFrame:
    liwc_analysis_per_keyword_mean = {}
    for keyword, liwc_analysis in liwc_analysis_per_keyword.items():
        liwc_analysis_mean = liwc_analysis.mean(axis=0).to_frame()
        liwc_analysis_per_keyword_mean[keyword] = liwc_analysis_mean.transpose()

    table = pd.DataFrame(
        pd.concat(list(liwc_analysis_per_keyword_mean.values())))
    table.insert(0, '', list(liwc_analysis_per_keyword_mean.keys()))
    table.set_index('', inplace=True)
    return table

# Faz o calculo do coeficiente de Gini.
def gini(x: list) -> float:
    total = 0
    for i, xi in enumerate(x[:-1], 1):
        total += np.sum(np.abs(xi - x[i:]))
    return total / (len(x)**2 * np.mean(x))

# Aplica o coeficiente de gini para pegar os atributos mais discriminantes
def apply_gini_coefficient_to_pick_the_most_discriminating_attributes(
    liwc_analysis_per_keyword: dict,
    selected_attributes: list = None,
    n_attribute: int = 20
) -> pd.DataFrame:
    scores_keyword = generate_table_with_average_analysis_liwc_by_keyword(
        liwc_analysis_per_keyword)
    if not selected_attributes:
        selected_attributes = scores_keyword.columns
    scores_keyword = scores_keyword.loc[:, selected_attributes]
    scores_keyword = scores_keyword.transpose()

    baseline = scores_keyword.mean(axis=1)
    scores_rel = scores_keyword.divide(baseline, axis=0)
    whitelist = scores_rel.apply(lambda s: gini(s.values), axis=1).sort_values(
        ascending=False).head(n_attribute).index
    scores_rel = scores_rel[scores_rel.index.isin(whitelist)]
    scores_rel.index = scores_rel.index.map(lambda s: s.title())
    scores_rel.rename(index=attribute_translation, inplace=True)

    scores_rel_zscore = (scores_rel - scores_rel.mean()) / scores_rel.std()
    return scores_rel_zscore

# Pega o top usuarios mais frequentes.
def get_top_frequent_users(

```

```

dataset: pd.DataFrame,
verified: bool,
is_retweet: bool,
n_users: int = 5
):
    if verified:
        filtered_users = dataset[dataset['verified'] == 'True']
    else:
        filtered_users = dataset[dataset['verified'] == 'False']

    if is_retweet:
        count_users = filtered_users['username_rt'].value_counts()
    else:
        count_users = filtered_users['screen_name'].value_counts()
    top_users = count_users.head(n_users)
    return top_users.to_dict()

# Pega a ocorrencia do usuario em cada palavra chave.
def get_user_occurrence_for_each_keyword(
    user: str,
    keywords_dir: str,
    is_retweet: bool
) -> dict:
    occurrences_per_keyword = {}
    # Define o tamanho maximo do campo em 100 MB.
    csv.field_size_limit(100000000)
    for keyword in os.listdir(keywords_dir):
        current_dir = keywords_dir + '/' + keyword + '/'
        current_dir_datasets = []
        for file in os.listdir(current_dir):
            if is_retweet and 'RT' in file:
                dataset = read_csv(current_dir + file)
                current_dir_datasets.append(dataset)
            elif not is_retweet and 'RT' not in file:
                dataset = read_csv(current_dir + file)
                current_dir_datasets.append(dataset)
        if len(current_dir_datasets) > 0:
            datasets = pd.concat(current_dir_datasets)
            if is_retweet:
                filtered_by_user = datasets[datasets['username_rt'] == user]
                count_user = filtered_by_user['username_rt'].value_counts()
            else:
                filtered_by_user = datasets[datasets['screen_name'] == user]
                count_user = filtered_by_user['screen_name'].value_counts()
            if len(count_user) > 0:
                occurrences_per_keyword[keyword] = count_user.values[0]
            else:
                occurrences_per_keyword[keyword] = 0
        else:
            occurrences_per_keyword[keyword] = 0
    return dict(sorted(occurrences_per_keyword.items(), key=lambda x: x[1], reverse=True))

# Pega o numero de seguidores do usuario.
def get_user_followers(user: str, dataset: pd.DataFrame, is_retweet: bool):
    if is_retweet:
        filtered_by_user = dataset[dataset['username_rt'] == user]
    else:
        filtered_by_user = dataset[dataset['screen_name'] == user]

```

```

    return filtered_by_user['followers'].max()

# Filtra o dataset com aqueles que possuem a palavra chave.
def get_data_with_the_keyword(dataset: pd.DataFrame, keyword: str) -> pd.DataFrame:
    return dataset[dataset['text'].str.contains(keyword, case=False)]

# Pega o sentimento de retweets por comunidade de cada palavra-chave.
def get_sentiment_retweet_per_communities_and_keywords(
    dataset_communities_dir: str,
    keywords: list = keywords_data,
    begin: int = 0,
    end: int = 4
) -> dict:
    communities = {'Comunidade_□' + str(i + 1): {}}
        for i in range(begin, end + 1)}
    # Define o tamanho maximo do campo em 100 MB.
    csv.field_size_limit(100000000)
    for keyword in keywords:
        for i in range(begin, end + 1):
            dataset_name = 'dataset_' + str(i) + '.csv'
            community = 'Comunidade_□' + str(i + 1)
            current_dataset_name = dataset_communities_dir + '/' + dataset_name
            dataset = read_csv(current_dataset_name)
            filtered_dataset = get_data_with_the_keyword(dataset, keyword)
            if filtered_dataset.shape[0] > 0:
                sentiment_polarity_column = filtered_dataset['sentiment_polarity'].astype(
                    float
                )
                sentiment_polarity = sentiment_polarity_column.mean()
                communities[community][keyword] = sentiment_polarity
            else:
                communities[community][keyword] = 0
    return communities

# Gera a analise LIWC das comunidades.
def generate_liwc_analysis_of_datasets_communities(
    dataset_communities_dir: str,
    begin: int = 0,
    end: int = 4
) -> dict:
    liwc_analysis = {}
    for i in range(begin, end + 1):
        dataset_name = 'dataset_' + str(i) + '.csv'
        community = 'Comunidade_□' + str(i + 1)
        current_dataset_name = dataset_communities_dir + '/' + dataset_name
        dataset = read_csv(current_dataset_name)
        liwc_dataset = liwc.process_df(dataset, 'text')
        liwc_analysis[community] = liwc_dataset
    return liwc_analysis

# Gera o grafico de distribuicao de retweets por tweet.
def generates_cdf_chart_of_retweet_occurrences(
    dataset: pd.DataFrame,
    title: str = '',
    xlabel: str = '',
    ylabel: str = ''

```

```

) -> None:
    # Conte a quantidade de ocorrencias para cada valor em 'tweet_id_rt'
    count = dataset['tweet_id_rt'].value_counts().reset_index()
    count.columns = ['tweet_id_rt', 'quantity']

    # Conte a quantidade de vezes que cada quantidade ocorre
    quantity_of_quantities = count['quantity'].value_counts().reset_index()
    quantity_of_quantities.columns = ['quantity', 'count']

    # Ordene o DataFrame por 'quantidade'
    quantity_of_quantities = quantity_of_quantities.sort_values(by='quantity')

    # Calcule a distribuicao acumulada de frequencia (CDF)
    cdf = np.cumsum(
        quantity_of_quantities['count']) / sum(quantity_of_quantities['count'])
    )

    # Plotagem do grafico de CDF
    plt.figure(figsize=(10, 6))
    plt.plot(quantity_of_quantities['quantity'], cdf, marker='o')
    plt.xscale('log')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.title(title)
    plt.grid(True)

    # Exiba o grafico
    plt.show()

def get_network_number_users(network_path: str) -> int:
    network = pd.read_csv(
        network_path,
        on_bad_lines='skip',
        dtype='unicode',
        delimiter='csv_limiter',
        engine="python",
        names=['Col1', 'Col2', 'edge_weight']
    )
    user_ids = network['Col1'].tolist() + network['Col2'].tolist()
    network_user_ids = pd.DataFrame({'user_id': user_ids})
    network_user_ids = network_user_ids.drop_duplicates().reset_index(drop=True)
    return network_user_ids.shape[0]

def get_partition(communities_dir: str, n_communities: int = None):
    files = os.listdir(communities_dir)
    files.remove('datasets')
    files = sorted(files, key=lambda x: int(x.split('_')[1]))
    partition = {}
    if n_communities != None:
        n_communities = len(files)
    for i, file in enumerate(files):
        if i == n_communities:
            break
        community_id = file.replace('.csv', '').split('_')[-1]
        community = nx.read_edgelist(
            communities_dir + '/' + file,
            delimiter='csv_limiter',
            data=[('weight', int)],

```

```

        create_using=nx.Graph
    )
    partition[community] = int(community_id)
return partition

def get_communities_dataframe(communities_dir: str):
    files = os.listdir(communities_dir)
    files.remove('datasets')
    files = sorted(files, key=lambda x: int(x.split('_')[1]))
    dataframes = {}
    for file in files:
        community_id = file.replace('.csv', '').split('_')[-1]
        community = pd.read_csv(
            communities_dir + '/' + file,
            on_bad_lines='skip',
            dtype='unicode',
            delimiter='csv_limiter',
            engine="python",
            names=['Col1', 'Col2']
        )
        dataframes[int(community_id)] = community
    return dataframes

# Pega o numero de retweets de cada comunidade,
def get_community_number_retweets(
    dataset: pd.DataFrame,
    community_user_ids: pd.DataFrame
) -> pd.DataFrame:
    user_rts = dataset[dataset['user_id'].isin(community_user_ids['user_id'])]
    return user_rts.shape[0]

# Pega os ids dos usuarios em um dataframe.
def get_community_users_ids(community: pd.DataFrame) -> pd.DataFrame:
    community.columns = ['Col1', 'Col2']
    user_ids = community['Col1'].tolist() + community['Col2'].tolist()
    community_user_ids = pd.DataFrame({'user_id': user_ids})
    return community_user_ids.drop_duplicates().reset_index(drop=True)

# Pega o numero de usuarios da comunidade.
def get_community_number_users(community_user_ids: pd.DataFrame) -> int:
    return community_user_ids.shape[0]

def generates_dataframe_with_community_info(
    dataset: pd.DataFrame,
    communities_dir: str
) -> pd.DataFrame:
    dataframes = get_communities_dataframe(communities_dir)
    df_data = {'community_id': [], 'community': [],
               'n_users': [], 'n_retweets': []}
    for community_id, community in dataframes.items():
        community_user_ids = get_community_users_ids(community)
        dataset_community = dataset[dataset['user_id'].isin(
            community_user_ids['user_id'])]
        n_users = get_community_number_users(community_user_ids)
        n_retweets = get_community_number_retweets(

```



```

        dataset_community, community_user_ids)
    df_data['community_id'].append(community_id)
    df_data['community'].append(community)
    df_data['n_users'].append(n_users)
    df_data['n_retweets'].append(n_retweets)
    return pd.DataFrame(df_data)

# Gera o grafico de distribuicao de usuarios por comunidade.
def generates_cdf_chart_of_users_per_community(
    communities_dir: str,
    dataset: pd.DataFrame,
    log_scale: bool = False,
    title: str = '',
    xlabel: str = '',
    ylabel: str = ''
) -> None:
    communities_info = generates_dataframe_with_community_info(
        dataset, communities_dir)
    # Ordene o DataFrame por 'n_users'.
    communities_info = communities_info.sort_values(by='n_users')

    # Calcule a distribuicao acumulada de frequencia (CDF)
    # cdf = np.cumsum(communities_info['n_users'] / n_users_network)
    cdf = np.cumsum(
        (communities_info['n_users'] * 0 + 1) / communities_info.shape[0])

    # Plotagem do grafico de CDF
    plt.figure(figsize=(10, 6))
    plt.plot(communities_info['n_users'], cdf, marker='o')
    if log_scale:
        plt.xscale('log')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.title(title)
    plt.grid(True)

    # Exiba o grafico
    plt.show()

# Gera grafico de numero de retweets por numero de vezes que um tweet foi retweetado.
def generates_graph_of_number_retweets_by_number_of_times_a_tweet_was_retweeted(
    dataset: pd.DataFrame
) -> None:
    # Conte a quantidade de ocorrencias para cada valor em 'tweet_id_rt'
    count = dataset['tweet_id_rt'].value_counts().reset_index()
    count.columns = ['tweet_id_rt', 'quantity']

    # Conte a quantidade de vezes que cada quantidade ocorre
    quantity_of_quantities = count['quantity'].value_counts().reset_index()
    quantity_of_quantities.columns = ['quantity', 'count']

    # Ordene o DataFrame por 'quantidade'
    quantity_of_quantities = quantity_of_quantities.sort_values(by='quantity')

    # Calcule a multiplicacao de "quantidade" e "contagem" para cada linha
    quantity_of_quantities['multiplication'] = quantity_of_quantities['quantity'] * \
        quantity_of_quantities['count']

```

```

# Crie o grafico de dispersao
plt.figure(figsize=(10, 6))
plt.plot(quantity_of_quantities['quantidade'],
         quantity_of_quantities['multiplication'], marker='o')
plt.xlabel('#Ocorrencias')
plt.ylabel('Quantidade de RTs')
plt.grid(True)

# Exiba o grafico
plt.show()

# Gera o dendrograma do grafo.
def generates_dendrogram(graph: nx.Graph) -> None:
    # Gera uma matriz de ligacao para o dendrograma.
    linkage_matrix = linkage(nx.convert_matrix.to_numpy_array(graph))

    dendrogram(linkage_matrix, labels=list(graph.nodes()), orientation='right')
    plt.show()

# Gera grafo.
def generates_graph(
    graph_path: str,
    k: float = 0.1,
    node_size: int = 2,
    node_color: str = 'skyblue',
    edges_width: float = 0.2,
    edges_alpha: int = 1,
    edges_color: str = 'gray'
) -> None:
    graph = nx.read_edgelist(
        graph_path,
        delimiter='csv_limiter',
        data=[('weight', int)],
        create_using=nx.Graph
    )

    # Layout do grafo (posicoes dos nos) com um valor de k ajustado
    pos = nx.spring_layout(graph, k=k)

    # Plotagem dos nos
    nx.draw(graph, pos, with_labels=False,
            node_size=node_size, node_color=node_color)

    # Plotagem das arestas
    nx.draw_networkx_edges(graph, pos, width=edges_width,
                           alpha=edges_alpha, edge_color=edges_color)

    # Exibir o grafico
    plt.show()

def get_graph_info(graph_path: str) -> None:
    graph = nx.read_edgelist(
        graph_path,
        delimiter='csv_limiter',
        data=[('weight', int)],
        create_using=nx.Graph
    )

```

```

number_of_nodes = graph.number_of_nodes()
print("Número de vértices:", number_of_nodes)
print("Número de arestas:", graph.number_of_edges())
print("Grau médio=", sum(dict(graph.degree()).values()) / number_of_nodes)
if nx.is_connected(graph):
    diameter = nx.diameter(graph)
else:
    diameter = '-'
print("Diâmetro=", diameter)

def get_tweets_with_that_word(
    dataset: pd.DataFrame,
    word: str,
    csv_result_name: str = 'tweets.csv'
):
    # Filtra o DataFrame para conter apenas linhas que
    # contem a palavra especifica na coluna 'text'
    filtered_dataset = dataset[dataset['text'].str.contains(word)][ 'text' ]
    filtered_dataset.to_csv(csv_result_name)

# Gera um grafo com a relacao passada como parametro.
def generate_users_related_graph(
    dataset: pd.DataFrame,
    related_column: str,
    csv_name: str,
    out_dir: str
) -> None:
    # Cria rede COMPLETA e adiciona arestas.
    edges = {}
    for _, row in tqdm.tqdm(dataset.iterrows()):
        source = row[related_column]
        target = row['user_id']

        if (source, target) in edges:
            edges[source, target] += 1
        else:
            edges[source, target] = 1

    # Cria a rede como grafo direcionado (DiGraph).
    edgelist = [(x[0], x[1], {"weight": edges[x]}) for x in edges]
    G = nx.DiGraph()
    G.add_edges_from(edgelist)

    if not (os.path.exists(out_dir)):
        os.mkdir(out_dir)

    # Escreve arquivo de saida.
    nx.write_edgelist(
        G,
        out_dir + '/' + csv_name,
        delimiter='csv_limiter',
        data=['weight'],
        encoding='utf-8'
    )

# Gera grafo de retweets em comum.
def generates_users_with_common_retweets_graph(

```

```

dataset: pd.DataFrame,
csv_name: str,
out_dir: str
) -> None:
    print("Dividindo DataFrame em lotes...")
    # Definir o tamanho do lote
    batch_size = 1000

    # Dividir o DataFrame em lotes
    data_batches = [dataset[i:i+batch_size] for i in range(0, len(dataset), batch_size)]

    print(f"Total de lotes: {len(data_batches)}")

    # Criar um grafo
    G = nx.Graph()

    # Processar cada lote
    for batch_index, batch in enumerate(data_batches):
        print(f"Processando lote {batch_index+1} de {len(data_batches)}")

        # Agrupar os dados pelo ID da postagem e coletar os
        # usuarios que compartilharam cada postagem
        grouped = batch.groupby('tweet_id_rt')['user_id'].apply(list).reset_index(
            name='user_id'
        )

        # Iterar pelos grupos e criar as arestas entre os
        # usuarios que compartilharam a mesma postagem
        for _, group in grouped.iterrows():
            user_ids = group['user_id']

            for i in range(len(user_ids)):
                for j in range(i + 1, len(user_ids)):
                    user1 = user_ids[i]
                    user2 = user_ids[j]

                    if G.has_edge(user1, user2):
                        G[user1][user2]['weight'] += 1
                    else:
                        G.add_edge(user1, user2, weight=1)

    if not (os.path.exists(out_dir)):
        os.mkdir(out_dir)

    # Escreve arquivo de saída.
    nx.write_edgelist(
        G,
        out_dir + '/' + csv_name,
        delimiter='csv_limiter',
        data=['weight'],
        encoding='utf-8'
    )

    print("Processamento concluído!")

# Pega os retweets.
def get_RTs(dataset: pd.DataFrame) -> None:
    return dataset[dataset['is_quote_status'] == 'False']

```

```

# Pega os quotes.
def get_quotes(dataset: pd.DataFrame) -> None:
    return dataset[dataset['is_quote_status'] == 'True']

# Pega os tweets que sao resposta.
def get_replys(dataset: pd.DataFrame) -> None:
    return dataset[dataset['in_reply_to_user_id'].notna()]

# Usuarios mais ativos (mais arestas chegando).
def net_in_degree(network: pd.DataFrame, users_dir: str, top_n: int = 5) -> pd.DataFrame:
    users = read_csv(users_dir)
    users['user_id'] = users['user_id'].astype(float)

    in_degree = dict(network.in_degree(weight='weight'))
    df_indeg = pd.DataFrame.from_dict(
        in_degree,
        orient='index',
        columns=['in_degree']
    ).reset_index().rename({'index': 'user_id'}, axis=1)
    df_indeg['user_id'] = df_indeg['user_id'].astype(float)

    return pd.merge(
        df_indeg,
        users,
        on='user_id',
        how='inner'
    ).sort_values(
        by='in_degree',
        ascending=False
    ).drop_duplicates(subset=['user_id'], keep='first').head(top_n)

# Usuarios mais influentes (mais arestas saindo).
def net_out_degree(network: pd.DataFrame, users_dir: str, top_n: int = 5) -> pd.DataFrame:
    users = read_csv(users_dir)
    users['user_id'] = users['user_id'].astype(float)

    out_degree = dict(network.out_degree(weight='weight'))
    df_outdeg = pd.DataFrame.from_dict(
        out_degree,
        orient='index',
        columns=['out_degree']
    ).reset_index().rename({'index': 'user_id'}, axis=1)
    df_outdeg['user_id'] = df_outdeg['user_id'].astype(float)

    return pd.merge(
        df_outdeg,
        users,
        on='user_id',
        how='inner'
    ).sort_values(
        by='out_degree',
        ascending=False
    ).drop_duplicates(subset=['user_id'], keep='first').head(top_n)

# Gera um grafo que liga os usuarios com retweets em comum.
def generates_a_graph_of_users_with_common_retweets(

```

```

    all_retweets_and_quotes: pd.DataFrame,
    per_week_dir: str = 'dataWorldCup/csv/weeks/separateTweetsAndRTs',
    graphs_dir: str = 'graphs/RTs',
    csv_name: str = 'relationship_graph_users_with_common_retweets_from_all_retweets.csv',
    per_week: bool = False
) -> None:
if not (os.path.exists(graphs_dir)):
    os.mkdir(graphs_dir)

    common_retweets_dir = graphs_dir + '/commonRetweets'
if not (os.path.exists(common_retweets_dir)):
    os.mkdir(common_retweets_dir)

    # Gera o grafo de todos os RTs.
    all_retweets = get_RT(all_retweets_and_quotes)
    generates_users_with_common_retweets_graph(
        dataset=all_retweets,
        csv_name=csv_name,
        out_dir=common_retweets_dir
    )

if per_week:
    generates_a_graph_of_users_with_common_retweets_per_week(per_week_dir, graphs_dir)

# Gera um grafo que liga os usuarios com retweets em comum.
def generates_a_graph_of_users_with_common_retweets_per_week(
    per_week_dir: str = 'dataWorldCup/csv/weeks/separateTweetsAndRTs',
    graphs_dir: str = 'graphs/RTs'
) -> None:
if not (os.path.exists(graphs_dir)):
    os.mkdir(graphs_dir)

    common_retweets_dir = graphs_dir + '/commonRetweets'
if not (os.path.exists(common_retweets_dir)):
    os.mkdir(common_retweets_dir)

    # Gera grafos por semanas dos RTs.
    common_retweets_per_week_dir = common_retweets_dir + '/perWeek'
if not (os.path.exists(common_retweets_per_week_dir)):
    os.mkdir(common_retweets_per_week_dir)
for filename in os.listdir(per_week_dir):
    if "_RT_" in filename:
        dataset_week = read_csv(per_week_dir + '/' + filename)
        week_RT = get_RT(dataset_week)
        csv_name = 'relationship_graph_users_with_common_retweets_from_' + filename
        generates_users_with_common_retweets_graph(
            dataset=week_RT,
            csv_name=csv_name,
            out_dir=common_retweets_per_week_dir
        )

# Gera grafo de relacao entre quem tweetou e quem retweetou.
def generates_a_graph_regarding_who_tweeted_and_who_retweeted(
    all_retweets_and_quotes: pd.DataFrame,
    csv_name: str = 'relationship_graph_tweeted_and_retweeted_from_all_retweets.csv',
    per_week_dir: str = 'dataWorldCup/csv/weeks/separateTweetsAndRTs',
    graphs_dir = 'graphs/RTs'
) -> None:
if not (os.path.exists(graphs_dir)):

```

```

    os.mkdir(graphs_dir)

# Gera o grafo de todos os RTs.
all_retweets = get_RTs(all_retweets_and_quotes)
tweet_retweet_dir = graphs_dir + '/relationshipTweetedAndRetweeted'
if not (os.path.exists(tweet_retweet_dir)):
    os.mkdir(tweet_retweet_dir)
generate_users_related_graph(
    dataset=all_retweets,
    related_column='user_id_rt',
    csv_name=csv_name,
    out_dir=tweet_retweet_dir
)

# Gera grafos por semanas dos RTs.
tweet_retweet_per_week_dir = tweet_retweet_dir + '/perWeek'
if not (os.path.exists(tweet_retweet_per_week_dir)):
    os.mkdir(tweet_retweet_per_week_dir)
for filename in os.listdir(per_week_dir):
    if "_RT_" in filename:
        dataset_week = read_csv(per_week_dir + '/' + filename)
        week_RTs = get_RTs(dataset_week)
        generate_users_related_graph(
            dataset=week_RTs,
            related_column='user_id_rt',
            csv_name='relationship_graph_tweeted_and_retweeted_from_' + filename,
            out_dir=tweet_retweet_per_week_dir
        )

# Gera grafo de relacao entre quem tweetou e quem citou.
def generates_a_graph_regarding_who_tweeted_and_who_quoted(
    all_retweets_and_quotes: pd.DataFrame,
    csv_name: str = 'relationship_graph_tweeted_and_quoted_from_all_quotes.csv',
    per_week_dir: str = 'dataWorldCup/csv/weeks/separateTweetsAndRTs',
    graphs_dir = 'graphs/quotes'
) -> None:
    if not (os.path.exists(graphs_dir)):
        os.mkdir(graphs_dir)

    all_quotes = get_quotes(all_retweets_and_quotes)
    generate_users_related_graph(
        dataset=all_quotes,
        related_column='user_id_rt',
        csv_name=csv_name,
        out_dir=graphs_dir
    )

# Gera grafos por semanas dos RTs.
quotes_per_week_dir = graphs_dir + '/perWeek'
if not (os.path.exists(quotes_per_week_dir)):
    os.mkdir(quotes_per_week_dir)
for filename in os.listdir(per_week_dir):
    if "_RT_" in filename:
        dataset_week = read_csv(per_week_dir + '/' + filename)
        week_quotes = get_quotes(dataset_week)
        generate_users_related_graph(
            dataset=week_quotes,
            related_column='user_id_rt',
            csv_name='relationship_graph_tweeted_and_quoted_from_' + filename,

```

```

        out_dir=quotes_per_week_dir
    )

# Gera grafo de relacao entre quem tweetou e quem respondeu.
def generates_a_graph_regarding_who_tweetou_and_who_replied(
    all_tweets: pd.DataFrame,
    csv_name: str = 'relationship_graph_tweetou_and_replied_from_all_replies.csv',
    per_week_dir: str = 'dataWorldCup/csv/weeks/separateTweetsAndRTs',
    graphs_dir = 'graphs/replies'
) -> None:
    if not (os.path.exists(graphs_dir)):
        os.mkdir(graphs_dir)

# Gera o grafo de todos os replies.
all_tweets = read_csv('dataWorldCup/csv/weeks/unifiedPerWeeks/dataset_info_0_8.csv')
all_replies = get_replies(dataset=all_tweets)
generate_users_related_graph(
    dataset=all_replies,
    related_column='in_reply_to_user_id',
    csv_name=csv_name,
    out_dir=graphs_dir
)

# Gera grafos por semanas dos replies.
replies_per_week_dir = graphs_dir + '/perWeek'
if not (os.path.exists(replies_per_week_dir)):
    os.mkdir(replies_per_week_dir)
for filename in os.listdir(per_week_dir):
    if "_RT_" not in filename:
        dataset_week = read_csv(per_week_dir + '/' + filename)
        week_replies = get_replies(dataset=dataset_week)
        csv_name = 'relationship_graph_tweetou_and_replied_from_' + filename
        generate_users_related_graph(
            dataset=week_replies,
            related_column='in_reply_to_user_id',
            csv_name=csv_name,
            out_dir=replies_per_week_dir
        )

# Pega as comunidades pela algoritmo de louvain.
def get_communities_by_louvain(
    graph: nx.Graph,
    is_directed: bool,
    out_dir: str,
    all_communities: bool = False,
    n_communities: int = 5
) -> None:
    # Crie um grafo nao direcionado, se necessario.
    if is_directed:
        graph = graph.to_undirected()

# Execute o algoritmo Louvain para deteccao de comunidades.
partition = community.best_partition(graph)
print('Algoritmo de louvain rodou.')
print('Modularidade =', community.modularity(partition, graph))

# Crie um dicionario para armazenar as comunidades e seus membros.
communities = {}

```



```

for node, community_id in partition.items():
    if community_id not in communities:
        communities[community_id] = []
    communities[community_id].append(node)

if not (os.path.exists(out_dir)):
    os.mkdir(out_dir)

# Ordene as comunidades com base no tamanho (numero de membros) em ordem decrescente
sorted_communities = sorted(communities.items(), key=lambda x: len(x[1]), reverse=True)

if all_communities:
    n_communities = len(sorted_communities)

# Salve cada comunidade em um arquivo CSV
for idx, (community_id, members) in enumerate(sorted_communities[:n_communities]):
    subgraph = graph.subgraph(members)
    nx.write_edgelist(
        subgraph,
        out_dir + '/' + f'community_{idx}_{community_id}.csv',
        delimiter='csv_limiter',
        data=['weight'],
        encoding='utf-8'
    )
    print(f'Comunidade_{idx}_foi_salva.')

# Pega as comunidades pela algoritmo de girvan newman.
def get_communities_by_girvan_newman(
    graph: nx.Graph,
    is_directed: bool,
    out_dir: str,
    all_communities: bool = False,
    n_communities: int = 5
) -> None:
    # Crie um grafo nao direcionado, se necessario.
    if is_directed:
        graph = graph.to_undirected()

    # Execute o algoritmo Girvan-Neuman para deteccao de comunidades.
    communities = next(girvan_newman(graph))

    # Converta as comunidades para uma lista para facilitar o processamento.
    communities = [list(community) for community in communities]

    if not os.path.exists(out_dir):
        os.mkdir(out_dir)

# Ordene as comunidades com base no tamanho (numero de membros) em ordem decrescente
sorted_communities = sorted(
    enumerate(communities),
    key=lambda x: len(x[1]),
    reverse=True
)

if all_communities:
    n_communities = len(sorted_communities)

# Salve cada comunidade em um arquivo CSV
for idx, (community_id, members) in enumerate(sorted_communities[:n_communities]):

```

```

subgraph = graph.subgraph(members)
nx.write_edgelist(
    subgraph,
    os.path.join(out_dir, f'community_{idx}_{community_id}.csv'),
    delimiter='csv_limiter',
    data=['weight'],
    encoding='utf-8'
)
print(f'Comunidade_{idx}_{community_id}.')

def get_communities_user_ids(community_csv: str) -> pd.DataFrame:
    df = pd.read_csv(
        community_csv,
        names=['Col1', 'Col2'],
        delimiter='csv_limiter',
        engine="python"
    )
    user_ids = df['Col1'].tolist() + df['Col2'].tolist()
    user_ids_df = pd.DataFrame({'user_id': user_ids})
    return user_ids_df.drop_duplicates().reset_index(drop=True)

def get_communities_datasets(all_data: pd.DataFrame, dir: str, n_communities: int = 10):
    all_data['user_id'] = all_data['user_id'].astype(int)

    if os.path.exists(dir):
        files = sorted(os.listdir(dir), key=lambda x: int(x.split('_')[1]))

    datasets_dir = dir + '/datasets'
    if not os.path.exists(datasets_dir):
        os.mkdir(datasets_dir)

    for i, file in enumerate(files):
        if i == n_communities:
            break
        user_ids_df = get_communities_user_ids(community_csv=dir + '/' + file)
        dataset = all_data[all_data['user_id'].isin(user_ids_df['user_id'])]
        csv_name = datasets_dir + '/dataset_' + str(i) + '.csv'
        dataset.to_csv(csv_name, index=False, sep='csv_limiter')

def get_only_retweets_with_a_frequency_greater_than_n(
    dataset: pd.DataFrame,
    n: int = 0
):
    # Calcule a contagem de ocorrências para cada 'tweet_id_rt'.
    count = dataset['tweet_id_rt'].value_counts().reset_index()
    count.columns = ['tweet_id_rt', 'occurrences']

    # Filtrar as linhas onde 'quantidade' eh maior ou igual a n.
    tweet_id_rt_frequent = count[count['occurrences'] >= n]

    # Use a funcao merge para manter apenas as linhas com 'tweet_id_rt' frequente.
    filtered = dataset.merge(tweet_id_rt_frequent, on='tweet_id_rt', how='inner')
    return filtered

def get_only_retweets_with_a_frequency_less_than_n(dataset: pd.DataFrame, n: int):
    # Calcule a contagem de ocorrências para cada 'tweet_id_rt'.

```

```
count = dataset['tweet_id_rt'].value_counts().reset_index()
count.columns = ['tweet_id_rt', 'occurrences']

# Filtrar as linhas onde 'quantidade' eh maior ou igual a n.
tweet_id_rt_frequent = count[count['occurrences'] < n]

# Use a funcao merge para manter apenas as linhas com 'tweet_id_rt' frequente.
filtered = dataset.merge(tweet_id_rt_frequent, on='tweet_id_rt', how='inner')
return filtered

def get_only_the_data_between(
    dataset: pd.DataFrame,
    begin: str = '2022-11-19',
    end: str = '2022-12-19'
):
    date_filter = (dataset['date'] >= begin) & (dataset['date'] <= end)
    return dataset[date_filter]
```

APÊNDICE B – ARTIGO SBC

Caracterização e análise de formação de comunidades no contexto da Copa do Mundo de 2022

Alisson Fabra da Silva¹, Carina Friedrich Dorneles¹, Ana Paula Couto da Silva²

¹Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brazil

²Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brazil

alissonfs100@gmail.com, carina.dorneles@ufsc.br, ana.coutosilva@dcc.ufmg.br

Abstract. *This study employs community detection algorithms on Twitter data collected before, during, and after the 2022 FIFA World Cup, aiming to understand user behavior and identify interest groups. Utilizing graphs to explore properties such as communities and individual influence, the analysis seeks to characterize and summarize the primary features of the collected data. The focus is to identify, analyze, and describe the formed communities, providing insights into user interactions and behavioral patterns in real-world situations.*

Resumo. *Este estudo emprega algoritmos de detecção de comunidades em dados do Twitter, coletados antes, durante e após a Copa do Mundo de 2022, com o intuito de compreender o comportamento dos usuários e identificar grupos de interesse. Utilizando grafos para explorar propriedades, como comunidades e a influência de indivíduos, a análise visa caracterizar e resumir as principais características dos dados coletados. O foco é identificar, analisar e descrever as comunidades formadas, fornecendo insights sobre interações e padrões de comportamento dos usuários em situações do mundo real.*

1. Introdução

As redes sociais são plataformas utilizadas pela população para expor seus pensamentos, encontrar pessoas que compartilham das mesmas ideias e gerar discussões. Pelo grande número de internautas e por sua utilização diária, os mais diversos temas são abordados, desde pequenos eventos do dia a dia a eventos maiores e polêmicos que envolvem um grande número de pessoas. Sendo assim, se torna um importante meio para a difusão de informações e pode ser utilizado com esse propósito, substituindo canais exclusivos de notícias ou jornais, por exemplo. Apesar de ser positivo para informar a população dos acontecimentos, também pode ser prejudicial por gerar grande alcance de informações falsas e polarizar discussões. Conforme o relatório da visão geral global, atualmente, a população total mundial é de 7,8 bilhões de pessoas, e destas, 4,2 bilhões são usuárias de redes sociais. Em um ano o número de pessoas que utilizam redes sociais aumentou 490 milhões, ou seja, um crescimento de 13%. Isto significa que 53% da população mundial utilizam redes sociais [Kemp 2021].

Ao longo dos últimos anos, a ampla utilização das redes sociais gerou um aumento na quantidade de influenciadores e impulsionou o conteúdo gerado por eles, sendo muitas

destas pessoas inicialmente não famosas. Como consequência, estas pessoas passaram a atrair uma multidão de seguidores, muitas vezes formando uma rede que compartilham das mesmas opiniões ou que debatem sobre elas. Essas redes alcançam muitos indivíduos tão rapidamente que, em pouco tempo, atingem âmbito nacional e até mesmo mundial. Estudá-las contribui para uma percepção maior de como a internet e as redes sociais, como um meio rápido e de fácil acessibilidade, facilitam essa influência dos formadores de opinião sobre a população. O Twitter é uma das 10 redes sociais mais populares no Brasil e uma das 15 mais utilizadas em todo o mundo [Volpato 2023], o que o torna um importante meio de difusão de informações e, conseqüentemente, uma ótima fonte de dados, permitindo a extração de opiniões de grande parte da população.

Nas redes sociais, interações tendem a ser agrupadas em comunidades [Coelho et al. 2013]. As comunidades são muito comuns em redes de informações, podendo existir várias. Os nodos dentro de uma comunidade costumam ser densamente conectados e podem se sobrepor, ou seja, participar de diversas comunidades. Isso ocorre porque em uma rede social pode existir interações com amigos, colegas, familiares e com pessoas aleatórios da plataforma, formando uma comunidade muito densa. As ligações entre comunidades são chamadas de pontes, responsáveis pelo pertencimento das comunidades a uma rede, diferenciando de uma ilha de nodos.

Existem diferentes algoritmos para detecção de comunidades em redes, como louvain, surprise e leiden [Girvan and Newman 2002]. A detecção de comunidade pode auxiliar mídias sociais a identificarem pessoas com interesses comuns e mantê-las firmemente conectadas. Também podem ser usadas no aprendizado de máquina para detectar grupos com propriedades semelhantes e extraí-los para, por exemplo, para identificar manipuladores dentro de uma rede social.

Este trabalho tem como objetivo analisar este fenômeno de disseminação de informações nas redes sociais, onde grandes redes de opiniões são formadas e as postagens de alto alcance atingem pessoas dos mais variados lugares, idades e classes sociais. O conjunto de dados utilizados neste trabalho foi disponibilizado pelo projeto "PRO-CORES: Caracterização e Modelagem de Processos de Contágio em Redes Sociais de Diferentes Domínios"¹ [Malagoli et al. 2021]. Os dados foram extraídos do Twitter, no qual o tema da coleta foi a Copa do Mundo de futebol da FIFA, que ocorreu em 2022 no Catar. Este é um evento de grande importância e paixão para grande parte dos brasileiros, pois desperta um intenso fervor patriótico e une o país em torno do esporte, uma vez que o futebol é profundamente enraizado na cultura brasileira, sendo considerado um símbolo de identidade nacional e orgulho [Penfold 2019]. A análise desses dados permite compreender as discussões, opiniões e diferentes perspectivas dos brasileiros sobre esse evento esportivo de grande impacto no país. Foram analisados quase 12 milhões de tweets coletados entre novembro e dezembro de 2022, cobrindo o período pré, durante e pós evento. A caracterização destes dados examina a divulgação de informações considerando dois aspectos cruciais: o envolvimento dos internautas e as propriedades dos conteúdos.

Além de explorar as conversas relacionadas ao futebol, este estudo também busca compreender como os brasileiros utilizam o Twitter para expressar opiniões, compartilha-

¹<https://procores.com.br/>

har informações e participar de discussões sobre temas relevantes relacionados à Copa do Mundo de 2022. Por meio do conceito de computação social, que estuda a interação entre pessoas e tecnologia, foi possível analisar a participação dos indivíduos nas redes sociais e a sua influência na disseminação de opiniões e informações durante o evento. Através dessa abordagem, foram analisadas e detectadas redes de opiniões por semelhança e aplicados algoritmos de detecção de comunidade.

1.1. Objetivos

Este trabalho tem como objetivo geral a caracterização e análise de formação de comunidades identificadas através de um algoritmo de detecção de comunidades no contexto da discussão dos brasileiros sobre a Copa do Mundo de 2022 no Twitter.

Tendo em vista este objetivo geral, os seguintes objetivos específicos:

- Caracterização do conjunto de dados;
- Análise exploratória do conjunto de dados, utilizando métodos como análise de nuvens de palavras, análise da popularidade das palavras chave, análise de sentimentos, análise psicolinguísticas etc;
- Modelagem, através de grafos, das postagens comuns entre os internautas. No contexto do Twitter, estas postagens comuns podem ser definidas através da re-postagem de um conteúdo exatamente igual, por exemplo;
- Uso de algoritmo de detecção de comunidades considerando os grafos modelados;
- Caracterização e comparação do conteúdo das comunidades detectadas.

2. Trabalhos Relacionados

Neste trabalho é abordada a identificação de comunidades em redes sociais, mais especificamente no Twitter. O objetivo principal é analisar como informações sobre a Copa do Mundo de 2022 foram difundidas entre os internautas a partir da análise de dados coletados dessa rede. Este capítulo apresenta alguns artigos relacionados às análises realizadas neste trabalho. Os artigos são divididos em três grupos principais: Redes de Opinião nas Redes Sociais, Análise de Dados do Twitter e Copas do Mundo.

2.1. Redes de Opinião em Redes Sociais

No estudo de Teixeira e Azevedo (2011), foram empregadas técnicas de Análise de Sentimentos para avaliar se informações coletadas do Facebook e Twitter poderiam ser utilizadas para prever valores comerciais de produtos ou serviços antes de seu lançamento no mercado. O crescente uso das redes sociais no cotidiano proporcionou novas oportunidades para análise de dados e contribuições ativas dos usuários, resultando em um vasto repositório de informações. As empresas reconheceram o valor dessas plataformas para promover seus produtos e analisar a percepção do público em relação a eles.

O trabalho de Abbade, Della Flora e Noro (2014) teve como objetivo analisar como estudantes universitários se comportam em relação à influência interpessoal nas redes sociais durante o processo de tomada de decisão de consumo. A pesquisa foi conduzida por meio de um levantamento (survey) com 200 estudantes de uma Instituição de Ensino Superior, onde foram avaliadas escalas para medir a disposição dos estudantes para influenciar e serem influenciados por contatos em redes sociais. Os resultados indicaram diferenças entre gêneros e mostraram que o tempo de acesso à internet tem uma correlação significativa com a propensão dos indivíduos a serem influenciados por seus contatos nas redes sociais.

2.2. Uso do Twitter para Discussão de Eventos Reais

No trabalho de Malagoli et al. (2021), foram analisados mais de 9 milhões de tweets em português sobre a vacinação contra a COVID-19 durante os estágios iniciais da campanha no Brasil e no mundo. Os resultados forneceram uma visão inicial da dinâmica do debate online sobre a vacinação, destacando como as pessoas utilizam o Twitter para compartilhar suas impressões e preocupações sobre o tema.

O estudo de Araujo et al. (2023) investigou a promoção coordenada de campanhas de propaganda política antecipadas no Twitter durante o período pré-eleitoral brasileiro de 2022. Utilizando modelagem de rede e análise de comunidades, os resultados revelaram a presença significativa de grupos promovendo conteúdo relacionado a diferentes pré-candidatos políticos, destacando diferenças entre as comunidades de direita e esquerda em termos de tamanho e volume de informações compartilhadas.

2.3. Redes Sociais e Edições da Copa do Mundo

O trabalho de Lins (2020) teve como objetivo identificar as variáveis que influenciam o comportamento de compra por impulso de acessórios de torcida durante megaeventos esportivos, como a Copa do Mundo de Futebol FIFA 2018. Os resultados destacaram a relação entre identidade nacional, fanatismo pelo evento e o impulso de compra, revelando que os homens tendem a demonstrar maior envolvimento com o evento em comparação às mulheres.

A pesquisa de Gastaldo (2009) realizou uma análise crítica sobre a abordagem midiática da Copa do Mundo no Brasil, explorando o papel da mídia na influência do interesse social pela competição. Utilizando dados de audiência e temas predominantes na televisão brasileira durante a Copa do Mundo de 1998, o estudo examinou como a mídia contribuiu para a construção do interesse coletivo pelo futebol no país.

3. Caracterização do Conjunto com Todos os Dados

Nesta seção foram apresentadas as principais análises considerando os quase 12 milhões de tweets/retweets coletados. Na seção 3.2 uma visão geral do conjunto de dados foi apresentada, enquanto na seção 3.3 os resultados das análises conduzidas foram detalhados, abrangendo análises de perfil dos indivíduos, emojis, sentimentos e aspectos da psicolinguística.

3.1. Tipos de Análises Realizadas

A caracterização dos dados é uma etapa baseada em análises realizadas para maior compreensão dos dados. Inicialmente, através de uma análise geral, foi verificado o volume de dados coletados ao longo do tempo, examinando nuvens de palavras e avaliando a popularidade das palavras-chave utilizadas na coleta. Além disso, foram realizadas análises de perfil dos indivíduos, de emojis, de sentimentos e psicolinguística.

A análise de perfil dos indivíduos foi examinada em relação a distribuição de tweets e retweets de acordo com o tipo de conta. Foram analisados os 10 emojis mais frequentes em tweets e retweets, fornecendo exemplos de como aparecem. Também foram feitas análises de sentimentos e psicolinguística.

3.2. Conjunto de Dados

O conjunto de dados utilizado foi obtido dentro do contexto do projeto "PROCORES: Caracterização e Modelagem de Processos de Contágio em Redes Sociais de Diferentes Domínios"². Foram coletados tweets que mencionam, ao menos uma vez, uma das palavras dentro deste conjunto de palavras-chave: *Argentina, BrasilNaCopa, Catar, CopaDoMundo2022, CopaDoMundoFIFA, CopaMundialFIFA, CopadoMundo, FIFA, FIFAWorldCup, Hexa, Messi, Neymar, Qatar2022, QatarWorldCup2022, RUMOAOHEXA, SeleccionBrasileira, Tite, neyday*. No total, foram coletados quase 12 milhões de tweets durante um período de 9 semanas. Essas semanas abrangeram o intervalo entre 01 de novembro e 31 de dezembro de 2022.

A Tabela 1 apresenta uma descrição inicial do conjunto de dados, mostrando os totais de tweets, retweets e internautas únicos por semana de coleta. Nas análises realizadas, considera-se tanto os tweets quanto os *replies*, que são respostas a tweets postados por outras pessoas. De maneira geral, observa-se uma tendência de crescimento no engajamento dos internautas, estimado pela frequência de postagens (tweets e retweets), na discussão dos tópicos relacionados à Copa do Mundo durante a atuação da Seleção Brasileira, sendo este da primeira à sexta semana. É possível observar que houve uma queda a partir da semana seguinte. Na oitava semana houve o crescimento de postagens e do engajamento dos internautas, uma vez que foi a semana da final do torneio.

Table 1. Principais estatísticas do conjunto de dados

Semana	Início	#Tweets	#Retweets	#Internautas Únicos
1	01-11-2022	45393	73125	89352
2	05-11-2022	258904	318535	317448
3	12-11-2022	213547	295782	285148
4	19-11-2022	1053277	1826730	998463
5	26-11-2022	1005601	1471705	883883
6	03-12-2022	902572	1346987	912662
7	10-12-2022	556430	740736	620022
8	17-12-2022	711005	1035255	717269
9	24-12-2022	221277	253299	269587

3.2.1. Análise Exploratória do Conjunto de Dados

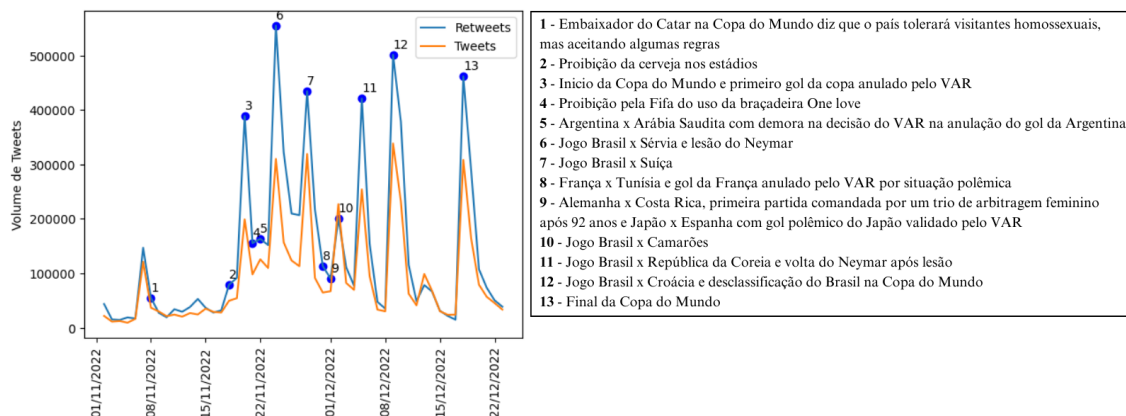
A Figura 1 apresenta uma análise mais detalhada da evolução temporal das discussões relacionadas à Copa do Mundo de 2022. Através da série temporal dos números diários de tweets e retweets, é possível observar a ocorrência de picos significativos que coincidem com eventos relevantes durante o período de coleta dos dados. Alguns desses eventos incluem os jogos da Seleção Brasileira e a proibição da cerveja nos estádios³.

Para concluir esta visão geral dos dados, a Figura 2 apresenta a popularidade das palavras-chave coletadas ao longo do tempo. A popularidade de uma palavra-chave em

²<https://procores.com.br/>

³<https://www.bbc.com/portuguese/internacional-63679803>

Figure 1. Volume de tweets durante o período coletado



uma determinada semana é proporcional ao número total de tweets ou retweets que mencionam essa palavra-chave nesse período. Essa análise foi representada através de mapas de calor, onde a cor de cada célula representa o total de tweets ou retweets que mencionam cada palavra-chave (eixo y) em uma determinada semana (eixo x), utilizando uma escala logarítmica.

A palavra-chave *Neymar* foi a mais mencionada. Sua popularidade foi mais alta durante a quarta semana até a sexta semana, que corresponde ao período em que a Seleção Brasileira jogou na Copa do Mundo. Durante esse tempo, Neymar, o jogador número 10 da equipe, sofreu uma lesão, se recuperou e voltou a jogar ainda durante a competição. Outra palavra-chave bastante citada nas semanas quatro e cinco foi *Catar*, que foi a sede do evento. Isso ocorreu porque essas semanas marcaram o início do torneio. *Tite* e *Hexa* também foram palavras bastante mencionadas. *Tite* é o apelido do técnico da seleção brasileira durante esse período, e *Hexa* refere-se à expectativa do Brasil de se tornar o único país hexacampeão da Copa do Mundo da FIFA. É interessante destacar que a palavra-chave *Messi* não registrou nenhuma menção durante a maior parte do período de coleta de dados, mas teve um aumento significativo de citações no final. Isso ocorreu durante a oitava semana, momento do torneio em que a Argentina, com Messi como jogador número 10, saiu vitoriosa. O mesmo ocorreu com a palavra *Argentina*.

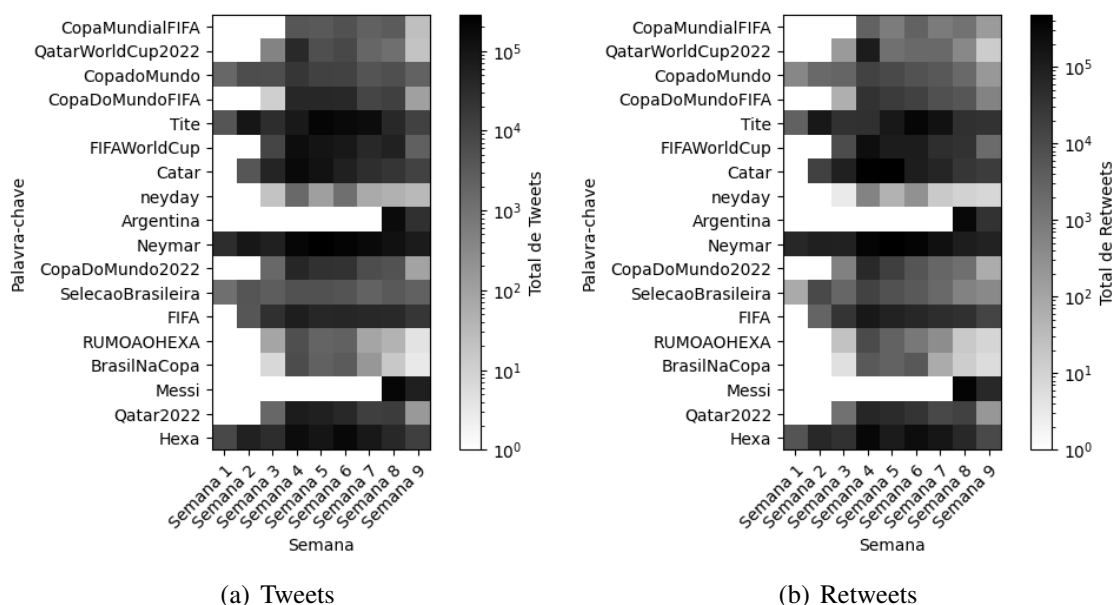
3.3. Resultados e Análises

Nesta seção, é detalhada e examinada uma série de análises conduzidas para compreender as diferentes perspectivas e discussões em torno da Copa do Mundo de 2022 no Twitter. Em cada análise, é descrita a metodologia utilizada e, em seguida, apresentada as principais conclusões obtidas.

3.3.1. Perfil dos Indivíduos

A análise do perfil dos indivíduos tem como propósito descrever as características destes que participaram ativamente das discussões sobre a Copa do Mundo de 2022 no Twitter. Até o final de 2022, as pessoas que utilizavam a plataforma Twitter eram divididos em duas categorias de contas: verificadas e não verificadas. A verificação da conta pelo

Figure 2. Popularidade das palavras-chave ao longo das semanas



Twitter indicava que essas pessoas despertam um maior interesse público e tendiam a se envolver mais nas conversas relacionadas a eventos significativos para a sociedade [Chen et al. 2020].

A Figura 3 ilustra as distribuições de probabilidade acumulada dos números de tweets e retweets feitos por pessoas com contas verificadas e não verificadas. Nos dados coletados as contas verificadas têm uma tendência maior a postar mais tweets e retweets. Cerca de 90% dos indivíduos com contas verificadas publicam até 30 tweets e 12 retweets, enquanto a mesma porcentagem com contas não verificadas publicam até 7 tweets e 9 retweets. O indivíduo mais ativo com uma conta verificada postou 7.930 tweets, enquanto o mais ativo com uma conta não verificada postou 28.084. Em relação aos retweets, verificou-se que as contas não verificadas mais ativas tendem a propagar mais informações, com um máximo de 381.894 retweets para uma única pessoa, em comparação com apenas 3.841 retweets da mais ativa com conta verificada.

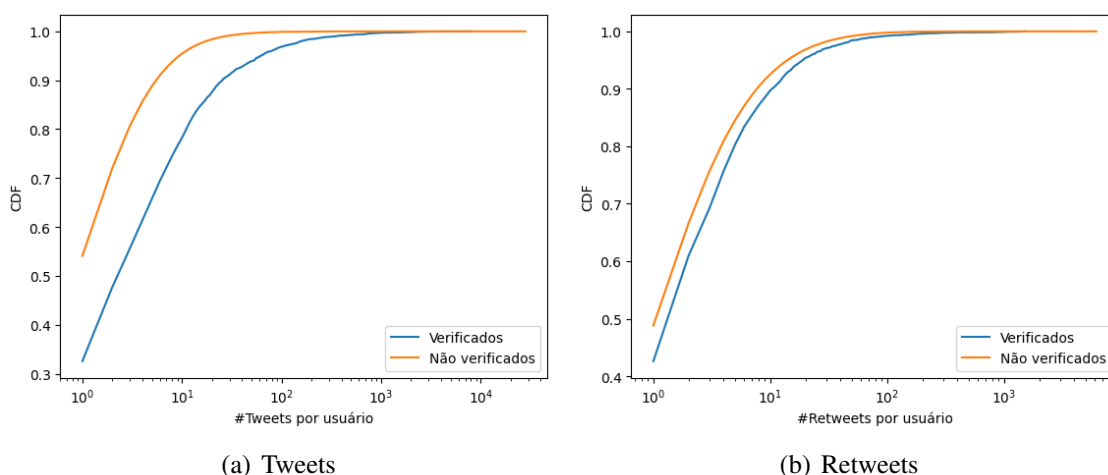
3.3.2. Análise dos Emojis

Os emojis são símbolos visuais usados nas redes sociais para complementar a comunicação por meio de texto. Eles representam emoções, ideias ou simbolismos. Segundo a análise dos dados, uma parte dos tweets (16,8%) e retweets (33,6%) continha pelo menos um emoji, identificados com o auxílio do pacote emoji⁴.

A Figura 4 apresenta os emojis mais utilizados entre os internautas que debatiam sobre a Copa do Mundo de 2022 no Twitter. Entre esses emojis que se destacaram, o emoji mais popular foi o do desenho da letra *R*, amplamente utilizado em tweets relacionados ao Brasil e à Argentina, representando as abreviações *BR* e *AR* para esses países. Os

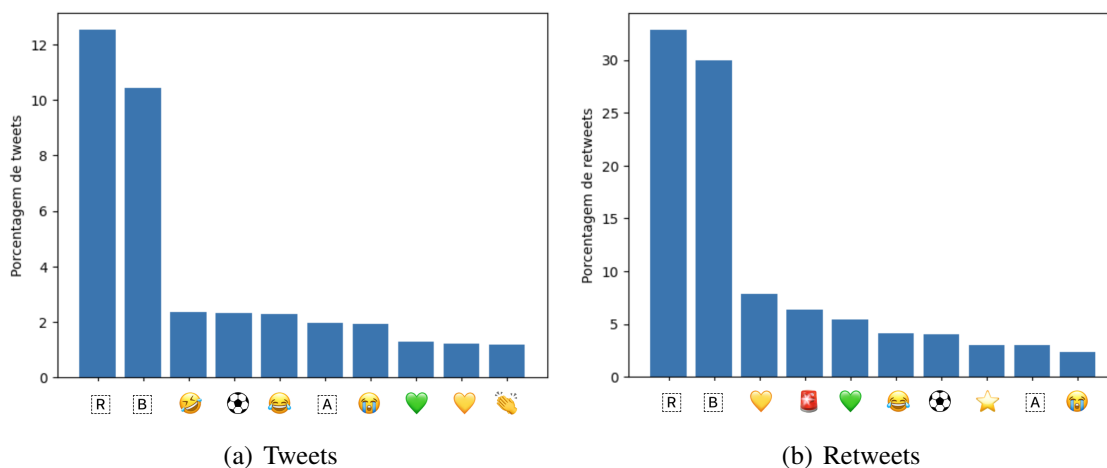
⁴<https://pypi.org/project/emoji/>

Figure 3. Distribuições dos números de tweets e retweets por tipo de conta



emojis de desenho das letras *B* e *A* também foram frequentemente usados em conjunto com o emoji do desenho da letra *R*. Outros dois emojis populares foram os corações nas cores verde e amarela, simbolizando o apoio à Seleção Brasileira no torneio. O emoji da bola de futebol também teve um uso significativo, o que era esperado dado o contexto da competição.

Figure 4. Top-10 emojis mais frequentes em tweets e retweets



Além dos emojis mencionados anteriormente, é notável a frequência elevada dos emojis de choro, que podem expressar tanto tristeza quanto felicidade. O choro de tristeza foi amplamente utilizado em decorrência da derrota do Brasil no torneio e das lesões sofridas por alguns jogadores brasileiros, que resultaram em desfalques na equipe. Por outro lado, os emojis de choro de felicidade, popularmente conhecidos como *chorar de rir*, foram bastante utilizados em diferentes contextos. Em alguns momentos, foram empregados para comentar situações engraçadas, enquanto em outros foram usados para expressar deboche.

Essa análise indica que os emojis desempenharam um papel importante na comunicação dos internautas no Twitter durante as discussões sobre a Copa do Mundo

de 2022, representando sentimentos, apoio, humor e ironia.

4. Análise e Detecção de Comunidades

Nesta seção foi feita a análise e detecção de comunidades. Dado o imenso volume de dados disponíveis, foram selecionados alguns dias específicos para a construção dessas redes, sendo o dia da cerimônia de abertura do evento, o dia da estreia da seleção brasileira, o dia da eliminação da seleção brasileira e o dia da partida final.

A rede gerada para rodar o algoritmo de detecção de comunidades foi uma rede baseada em retweets comuns entre pessoas. Estas redes foram representadas como grafos não direcionados, nos quais os vértices correspondem as pessoas e as arestas indicam que elas retweetaram o mesmo conteúdo. O peso das arestas representa a quantidade de retweets em comum entre esses indivíduos no conjunto de dados. Portanto, se o peso de uma aresta for, por exemplo, 10, isso significa que essas pessoas retweetaram 10 tweets idênticos, ou seja, com o mesmo ID, não apenas o mesmo texto.

A finalidade de representar o grafo dessa maneira foi criar uma rede de opinião, na qual a conexão entre os internautas é mais forte quanto mais informações compartilhadas eles têm em comum, como retweets do mesmo conteúdo. Nestas redes o algoritmo de detecção de comunidade de Louvain foi utilizado para obter as comunidades existentes neste grafo. Por conta de limitação computacional, este algoritmo de detecção de comunidades foi escolhido por ser eficiente [Aires and Nakamura 2017].

A qualidade das comunidades obtidas pelo algoritmo de Louvain foi avaliada usando a métrica de modularidade, que indica o grau de conexão entre os elementos de uma comunidade. Essa métrica varia de -1 a 1 na implementação do algoritmo de Louvain utilizada. Nos dias selecionados, foram observadas modularidades próximas a 1, o que sugere uma detecção eficaz de comunidades, pois, nessa implementação, valores mais próximos de 1 indicam uma detecção de alta qualidade.

Foram realizadas algumas análises para cada uma das 5 maiores comunidades dos dias selecionados. Para cada um dos dias está sendo ilustrado um tipo de análise diferente, para que dessa forma seja possível verificar a importância de cada uma delas e como elas são aplicadas. Entre as análises feitas estão análises de nuvens de pares de palavras, de sentimentos, psicolinguística e de tópicos LDA.

4.1. Dia da Cerimônia de abertura da Copa do Mundo de 2022

A rede gerada para o dia da cerimônia de abertura contava com 164.281 internautas e revelou a existência de 2.151 comunidades. A Tabela 2 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades. As comunidades encontradas apresentaram uma modularidade de 0,70.

Neste dia será mostrada a análise de nuvens dos 50 pares de palavras mais frequentes, nessas nuvens quanto maior o par de palavra mais frequente ele é. Na Figura 5 estão disponíveis as nuvens de pares de palavras para as 5 maiores comunidades. Na comunidade 1, a maior comunidade em número de internautas, os pares de palavras que mais apareceram foram *catar qatarworldcup2022*, *mundu qatarworldcup2022* e *jungkook bts*. Essas parecem estar mais ligadas ao início do evento de modo geral. Jungkook é um

Table 2. Informações gerais sobre os grafos do dia da cerimônia de abertura

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	164.281	12.055.873	146,77	-
Comunidade 1	27.350	2.493.228	182,32	7
Comunidade 2	23.425	770.294	65,77	9
Comunidade 3	15.253	592.159	77,64	10
Comunidade 4	8.278	261.549	63,191	9
Comunidade 5	8.077	243.416	60,27	10

integrante da banda BTS, que se apresentou na cerimônia de abertura⁵. Além disso as palavras *catar* e *qatarworldcup2022* fazem parte das palavras chaves utilizadas na coleta dos dados. As comunidades 2, 4 e 5 destacaram pares de palavras semelhantes, apareceram em todas *país anfitrião*, *direitos humanos* e *primeiro país*. Discussões sobre os direitos humanos no Catar, país sede da Copa do Mundo de 2022, foram levantadas no período do início do evento⁶, o que pode ter gerado tanta recorrência dos pares de palavras citados. A nuvem gerada para a comunidade 3 está diferente das demais, ela está com pares de palavras de tamanho muito uniforme e com uma mistura das que já apareceram nas outras nuvens.

4.2. Dia da Estreia da Seleção Brasileira na Competição

A rede gerada para o dia da estreia da seleção brasileira na competição incluía 266.622 internautas e resultou em 3.527 comunidades. A Tabela 3 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades. As comunidades encontradas apresentaram uma modularidade de 0,72.

Table 3. Informações gerais sobre os grafos do dia da estreia da seleção brasileira

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	266.622	23.252.970	174,43	-
Comunidade 1	21.209	506.842	47,79	12
Comunidade 2	15.881	1.647.442	207,47	8
Comunidade 3	13.445	1.336.024	198,74	8
Comunidade 4	8.204	496.100	120,94	8
Comunidade 5	7.281	342.204	93	11

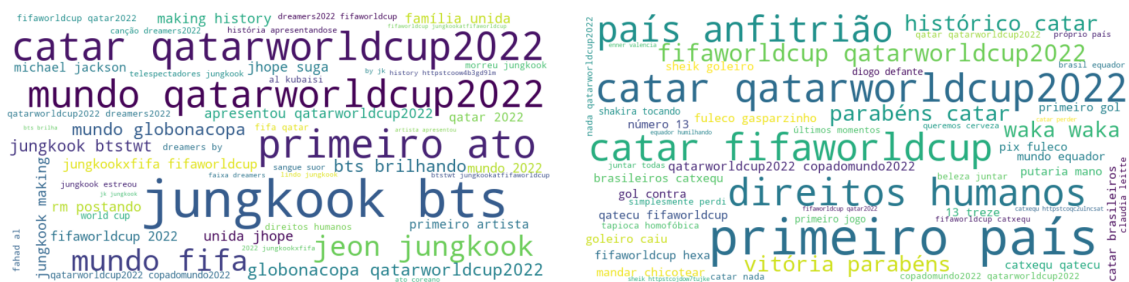
Para este dia será mostrada a análise de sentimentos. Nessa análise foi utilizada a ferramenta *LeIA*⁷, que extrai o sentimento de cada tweet. O *LeIA* fornece uma pontuação inteira que descreve o sentimento predominante no texto. Foram considerados tweets e retweets com pontuações menores que -0,05 como negativos, maiores que 0,05 como positivos e entre -0,05 e 0,05 como neutros. O *LeIA* é uma ferramenta muito utilizada

⁵<https://ge.globo.com/futebol/copa-do-mundo/noticia/2022/11/20/bts-na-copa-do-mundo-do-catar-jungkook-na-abertura-agita-fas-de-k-pop.ghtml>

⁶<https://www.cnnbrasil.com.br/esportes/copa-do-mundo-entenda-as-denuncias-sobre-dir>

⁷<https://github.com/rafjaa/LeIA>

Figure 5. Nuvem dos 50 pares de palavras mais populares das 5 maiores comunidades do dia da cerimônia de abertura



(a) Comunidade 1

(b) Comunidade 2



(c) Comunidade 3



(d) Comunidade 4



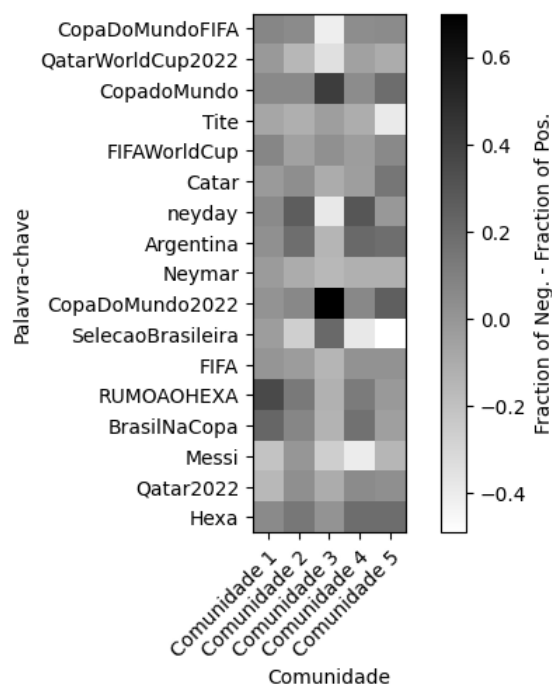
(e) Comunidade 5

para análise de sentimentos em textos de redes sociais, mas ele também pode ser aplicado a textos de outros domínios. A fim de entender como o sentimento expresso pelas pessoas varia, foi utilizado um mapa de calor para mostrar o contraste dos sentimentos expressos nos tweets e retweets através da pontuação contrastiva, sendo calculada como a diferença entre a fração de tweets positivos e negativos.

A Figura 6 apresenta a análise de sentimentos para este dia. Dentre os dias analisados, esse foi o que apresentou mais sentimentos negativos. A comunidade 1, a maior em número de internautas, apresenta sentimentos positivos para as palavras-chave *RUMAOHEXA* e *BrasilNaCopa*, demonstrando um ânimo relacionado a estreia da seleção brasileira. As comunidades 2, 4 e 5 tiveram sentimento negativo relacionado às palavras *SelecaoBrasileira* e *Tite*, o que pode significar uma possível insatisfação a respeito do jogo de estreia. A comunidade 3 obteve sentimento mais negativo para palavras-chave relacionadas ao Brasil na copa, como *RUMAOHEXA*, *BrasilNaCopa*, *Neymar* e *neyday*.

Além de apresentar sentimentos bastante positivos e negativos relacionados ao evento de modo geral, com *CopaDoMundoFIFA* e *QatarWorldCup2022* com sentimentos muito negativos e *CopaDoMundo2022* com sentimento muito positivo. Diferente do dia da cerimônia de abertura, a palavra *Messi* apareceu com sentimento mais negativo nessas comunidades.

Figure 6. Análise de sentimento por palavras chave das top 5 maiores comunidades do dia da estreia da seleção brasileira



4.3. Dia da Eliminação da Seleção Brasileira na Competição

A rede gerada para o dia da eliminação da seleção brasileira abrange 226.491 internautas e revelou 4.011 comunidades. A Tabela 4 mostra os dados gerais dos grafos da rede completa e das 5 maiores comunidades. As comunidades identificadas exibiram uma modularidade de 0,82.

Table 4. Informações gerais sobre os grafos do dia da eliminação da seleção brasileira

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	226.491	14.123.676	124,72	-
Comunidade 1	28.532	643.154	45,08	11
Comunidade 2	24.877	466.058	37,47	10
Comunidade 3	15.769	344.002	43,63	12
Comunidade 4	15.313	271.244	35,43	13
Comunidade 5	14.135	243.236	34,42	11

A análise escolhida para este dia é a análise psicolinguística. Para ela foi utilizado o léxico LIWC⁸ [Tausczik and Pennebaker 2010] para categorizar as palavras em

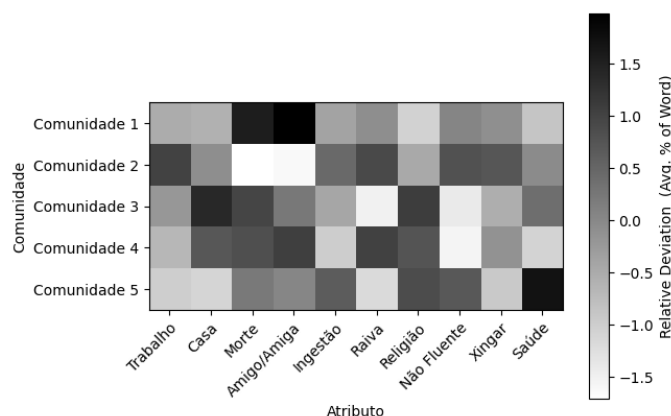
⁸<http://143.107.183.175:21380/portlex/index.php/pt/projetos/liwc>

diferentes atributos relacionados ao estilo linguístico, conceitos afetivos e cognitivos. A frequência média desses atributos foi calculada para cada palavra-chave nos tweets e retweets. Nos tweets e retweets analisados, foram identificados a presença dos 64 atributos disponíveis na versão em português do LIWC de 2007. Em seguida, foram analisadas diferenças estatísticas entre os debates em torno de diferentes palavras-chave, explorando a frequência média dos atributos nos tweets e retweets associados a cada palavra-chave. O teste não paramétrico de Kruskal-Wallis [Kruskal and Wallis 1952] foi utilizado para selecionar atributos que apresentassem diferenças significativas entre as palavras-chave. No entanto, foi observado que todos os 64 atributos apresentaram diferenças significativas.

Para lidar com a grande quantidade de atributos, o coeficiente de Gini [Yitzhaki 1979] foi utilizado para selecionar, dentre os 64, os 20 que eram mais discriminantes. Dessa forma, foi possível identificar os atributos mais relevantes para cada palavra-chave através de um mapa de calor com esses atributos discriminantes, considerando todos os tweets e retweets. Cada célula do mapa de calor em uma coluna representa o desvio relativo de um determinado atributo para uma palavra-chave específica em relação às outras palavras-chave. As células foram coloridas em gradiente entre vermelho e azul, indicando se o atributo está acima ou abaixo da média, respectivamente. Para isso, cada coluna foi normalizada utilizando a métrica z-score, ou seja, subtraímos a média da coluna e dividimos pelo desvio padrão da coluna

A Figura 7 apresenta a análise psicolinguística. Na comunidade 1 as palavras estão predominantemente ligadas a temas de amizade e morte. Na comunidade 2 há uma forte associação com palavras relacionadas a trabalho, raiva, xingamentos e não fluência. Nas comunidades 3 e 4 os retweets se concentram em discussões sobre casa, morte e religião, com a comunidade 4 também incluindo amizade e raiva em suas postagens. Por fim, a comunidade 5 se destaca por conter muitas palavras relacionadas à religião, dificuldades de comunicação e, principalmente, questões de saúde.

Figure 7. Análise de psicolinguística das top 5 maiores comunidades do dia da eliminação da seleção brasileira



4.4. Dia da Partida Final da Copa do Mundo de 2022

A rede gerada para esse dia da final da competição contou com 193.994 internautas e resultou em 4.298 comunidades. A Tabela 5 mostra os dados gerais dos grafos da rede

Table 5. Informações gerais sobre os grafos do dia da partida final

	#Vértices	#Arestas	Grau médio	Diâmetro
Rede completa	193.994	11.816.209	121,82	-
Comunidade 1	13.987	442.849	63,32	11
Comunidade 2	10.762	285.500	53,06	11
Comunidade 3	8.679	315.032	72,6	11
Comunidade 4	8.454	180.391	42,68	12
Comunidade 5	8.294	190.430	45,92	11

completa e das 5 maiores comunidades. As comunidades encontradas apresentaram uma modularidade de 0,79.

Para o dia da final foi escolhida a análise de tópicos LDA. Essa análise foi feita utilizando uma implementação do algoritmo LDA em python⁹. Foram analisados os retweets de cada uma das comunidades que foram trabalhadas. Antes de gerar a análise foi necessário verificar a quantidade de tópicos ideal de se trabalhar, isso foi feito verificando a coerência que cada quantidade de tópicos tinha. Por conta disso, algumas comunidades possuem menos tópicos nas análises.

Na tabela 6 está a análise de tópicos para a comunidade 1 do dia da partida final do evento. Nela foram identificados 8 tópicos, sendo que a maioria engloba tweets que comentam sobre o jogo e parabenizam a seleção argentina e o jogador Messi por ter conquistado o torneio. No tópico 2 aparece a palavra *fome*, que é por conta de tweets que diziam que a seleção argentina merecia ganhar o torneio porque a situação econômica estava ruim e o povo estava passando fome. No tópico 6 existem alguns tweets comparando a escolha do técnico da seleção argentina de colocar o melhor jogador do time nesse aspecto para bater o primeiro pênalti das disputas, com a escolha do técnico da seleção brasileira que colocou o melhor jogador para bater o último pênalti. Além disso, existem tweets que se referem a uma fala que o jogador francês Mbappé fez um tempo antes da Copa do Mundo de 2022 falando que o futebol sul-americano não é tão avançado quanto o da Europa¹⁰. No tópico 7, além de falar a respeito do jogo, também houveram tweets que falaram a respeito de manifestações a favor do governo Bolsonaro após a vitória de Lula nas eleições que ocorriam no período do evento¹¹. A palavra *patriotas* neste tópico representa isso.

A Tabela 7 mostra a análise para a comunidade 2. Nesta comunidade o número de tópicos identificados foi 10. Os tópicos 1, 2, 4, 8, 9 e 10 abordam o desempenho do jogador Messi no torneio e a conquista do título mundial pela Argentina. O tópico 3 fala a respeito do desempenho dos jogadores Neymar, Messi e Mbappé, que atuavam juntos no time Paris Saint-Germain Football Club. Os tópicos 5, 6 e 7 possuem críticas a escolha do técnico da seleção brasileira por não ter colocado o jogador principal desta função do time para bater primeiro na disputa de pênaltis como foi com a França e a Argentina.

Na Tabela 8 disponível a análise da comunidade 3, com 10 tópicos identificados.

⁹<https://pypi.org/project/gensim/>

¹⁰<https://www.cnnbrasil.com.br/esportes/em-maio-mbappe-disse-que-futebol-sul-americano>

¹¹<https://www.metropoles.com/colunas/grande-angular/patriotas-convocam-ultimo-ato-apos-45-dias-sem-resposta>

Table 6. Análise de tópicos LDA para a comunidade 1 do dia da partida final

Tópico 1	campeão	messi	argentina	cerimedo	fernando
Tópico 2	povo	argentina	merece	fome	jogador
Tópico 3	argentina	técnico	globo	vida	messi
Tópico 4	mundo	parabéns	copa	argentina	chegar
Tópico 5	história	seleção	comando	pessoas	brasil
Tópico 6	tite	aprende	messi	mbappé	argentina
Tópico 7	primeiro	pênalti	messi	defesa	patriotas
Tópico 8	messi	ganhou	garra	série	netflix

Table 7. Análise de tópicos LDA para a comunidade 2 do dia da partida final

Tópico 1	melhor	copa	messi	jogador	mundo
Tópico 2	messi	final	mundo	melhor	troféu
Tópico 3	neymar	messi	mbappé	final	copa
Tópico 4	messi	copa	mundo	futebol	lionel
Tópico 5	tite	messi	aprende	título	argentina
Tópico 6	tite	messi	lionel	campeão	seleção
Tópico 7	neymar	aprendeu	messi	copa	jogo
Tópico 8	contra	gol	arg	gols	austrália
Tópico 9	scaloni	primeiro	time	técnico	acordo
Tópico 10	bater	messi	sonho	abriu	partida

O tópico 1 fala sobre a situação polêmica já comentada anteriormente a respeito da escolha do técnico da seleção brasileira na escolha da ordem dos jogadores na disputa de pênaltis. O tópico 2 comenta a respeito do desempenho dos jogadores Neymar, Mbappé e Messi no torneio. A discussão do tópico 3 gira em torno de uma fala do comentarista brasileiro Casagrande que diz que o futebol brasileiro está atrasado se comparado com cenário mundial. O tópico 4 se refere ao desentendimento que o jogador Neymar teve com a influenciadora digital Nath Finaças após ela questionar o Neymar sobre declaração de imposto de renda. Os outros tópicos abordam comentários a respeito das conquistas do jogador Messi e também da seleção argentina.

Table 8. Análise de tópicos LDA para a comunidade 3 do dia da partida final

Tópico 1	tite	copa	mundo	neymar	mbappé
Tópico 2	neymar	messi	instagram	mbappé	qatar2022
Tópico 3	primeiro	casagrande	messi	argentina	frança
Tópico 4	copa	finanças	nath	twitter	neymar
Tópico 5	gaveta	hexa	copa	ouro	taça
Tópico 6	time	copa	live	messi	35
Tópico 7	psg	campeões	mbappé	jr	messi
Tópico 8	chegando	pegar	argentina	copa	futebol
Tópico 9	times	cabo	mbappé	copa	lacrando
Tópico 10	dessa	simplesmente	precisava	sempre	odeiam

As comunidades 4 e 5 obtiveram 8 e 10 tópicos respectivamente. Nessas comunidades foram discutidos temas já comentados nas outras comunidades, como a conquista

do título mundial pela seleção argentina, desempenho dos jogadores Neymar, Mbappé e Messi e polêmica da escolha do da ordem de jogadores pelo técnico da seleção brasileira nos pênaltis.

5. Considerações Finais e Trabalhos Futuros

Neste trabalho com o objetivo de realizar a identificação e análise de redes de opinião a partir de algoritmos de detecção de comunidades, foi realizada uma etapa de caracterização do conjunto de dados para que fossem aplicados esses algoritmos de maneira correta e mais eficiente. Desse modo, foram explorados modelos de análises, como a de evolução do volume de dados, a de popularidade das palavras chaves em cada semana, a de perfil dos indivíduos e de emojis.

Após a etapa inicial de caracterização do conjunto total dos dados, foi realizada a análise e detecção de comunidades a partir de redes de internautas com retweets comuns geradas. Foram encontradas limitações computacionais que dificultaram este processo. Por conta disso, a abordagem utilizada para contornar essa situação foi a escolha de dias marcantes durante este período para que fosse possível desenvolver o trabalho. Os dias escolhidos foram: o dia da cerimônia de abertura, da estreia da seleção brasileira na competição, da eliminação da seleção brasileira e da partida final da competição.

Com as comunidades já detectadas pelo algoritmo de detecção de comunidades de Louvain, foram escolhidas as cinco maiores comunidades em número de contas para fazer a análise. Dessa forma, na análise dessas comunidades algumas ferramentas como nuvens de pares de palavras, LeIA, LIWC, teste de Kruskal-Wallis, coeficiente de Gini e algoritmo LDA.

Este trabalho mostrou detalhadamente como fazer algumas análises para fazer uma caracterização de dados. Cada uma das análises aplicadas a cada uma das comunidades gerou uma perspectiva diferente a respeito do que estava sendo discutido. Nas nuvens de palavras conjugadas foi possível observar mais facilmente os assuntos debatidos nos dias selecionados. Já nas análises de sentimento, é possível ter uma perspectiva maior a respeito dos sentimentos aplicados nas palavras escolhidas para as postagens. Na análise psicolinguística também é possível ter uma noção dos sentimentos nas palavras, mas também podemos observar as palavras categorizadas em diferentes atributos relacionados ao estilo linguístico, conceitos afetivos e cognitivos. Com a análise de tópicos LDA podem ser identificados diversos temas que foram discutidos no conjunto de dados.

A análise realizada neste trabalho permite discernir como os eventos externos exercem influência significativa nas interações online, criando extensas redes de indivíduos alinhados em suas opiniões. Tal compreensão reforça a constatação de que a internet não apenas reflete, mas também impulsiona a propagação de informações. Essa constatação destaca seu papel não apenas como um espelho dos acontecimentos, mas como um substituto vital para os meios tradicionais de comunicação, como jornais e canais de notícias especializados. Assim, é possível compreender sobre a interconexão entre eventos globais e o ambiente virtual, evidenciando a relevância deste estudo para a compreensão do impacto das dinâmicas online na formação de opinião.

No que se refere a trabalhos futuros, são propostas as seguintes ideias:

- Geração da rede do conjunto de dados completo;

- Extração de backbone da rede com todo o conjunto de dados;
- Detecção de comunidades do conjunto de dados completo;
- Comparação de comunidades detectadas com diferentes algoritmos;
- Análise das comunidades detectadas.

References

- Aires, V. and Nakamura, F. (2017). Detecção de comunidades em redes sociais: Relacionando o método louvain a medidas de centralidade. In *Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC*, Porto Alegre, RS, Brasil. SBC.
- Belegante, T. C. and Menezes, L. P. (2015). A influência dos formadores de opinião nas redes sociais. *Anais do 11º ENCITEC 2015*. Acesso em: 12/06/2023.
- Chen, E., Lerman, K., and Ferrara, E. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273.
- Coelho, M. A. N., Borges, C. C. H., Neto, R. F., Vieira, A. B., and da Silva, A. P. C. (2013). Estratégia online para predição estruturada em redes complexas. In Braga, A. d. P. and Bastos Filho, C. J. A., editors, *Anais do 11 Congresso Brasileiro de Inteligência Computacional*, pages 1–6, Porto de Galinhas, PE. SBIC.
- Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.
- Jayawickrama, T. D. (2021). Community detection algorithms. *Towards Data Science*. Acesso em: 12/06/2023.
- Kemp, S. (2021). Digital 2021: global overview report. Acesso em: 27/11/2023.
- Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.
- Lins, S. (2020). Preparando-se para a copa do mundo: o que leva os brasileiros a comprar impulsivamente produtos para apoiar o seu país? In *Atas do X Simpósio Nacional de Investigação em Psicologia*. FPCEUP.
- Malagoli, L., Stancioli, J., Ferreira, C., Vasconcelos, M., Silva, A. P., and Almeida, J. (2021). Caracterização do debate no twitter sobre a vacinação contra a covid-19 no brasil. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 55–66, Porto Alegre, RS, Brasil. SBC.
- Penfold, T. (2019). National identity and sporting mega-events in brazil. *Sport in Society*, 22(3):384–398.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Volpato, B. (2023). Ranking: as redes sociais mais usadas no brasil e no mundo em 2023, com insights, ferramentas e materiais. Acesso em: 12/06/2023.
- Yitzhaki, S. (1979). Relative deprivation and the gini coefficient. *The Quarterly Journal of Economics*, 93(2):321–324.