Caracterização de dados do Twitter na Copa do Mundo'22

Alisson F. da Silva[1], **Carina F. Dorneles**[1], **Ana Paula Couto da Silva**[2]

[1]Universidade Federal de Santa Catarina
Instituto de Informática e Estatística – Florianópolis – SC – Brasil

[2]Universidade Federal de Minas Gerais
Departamento de Ciência da Computação – Minas Gerais – MG – Brasil

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

Abstract. *X, formerly known as Twitter, is a leading social media platform widely used for discussing diverse topics. During the 2022 World Cup, it gained immense popularity, particularly among Brazilian soccer enthusiasts. This article analyzes about 12 million tweets spanning before, during, and after the event, shedding light on how people utilize the platform to express their opinions and interests online.*

Resumo. *O X, previamente conhecido como Twitter, é uma das plataformas sociais mais populares, amplamente usada para discutir diversos temas. Durante a Copa do Mundo de 2022, foi especialmente movimentado pelos brasileiros devido ao seu entusiasmo pelo futebol. Este artigo analisa cerca de 12 milhões de tweets coletados antes, durante e após o evento, contribuindo para entender como as pessoas usam a internet para expressar suas opiniões e interesses.*

1. Introdução

As redes sociais são plataformas amplamente utilizadas para compartilhar pensamentos, conectar pessoas com interesses semelhantes e gerar debates sobre uma variedade de temas, desde assuntos cotidianos até questões controversas que envolvem muitas pessoas. Elas desempenham um papel crucial na disseminação de informações, muitas vezes substituindo canais tradicionais de notícias. No entanto, também enfrentam desafios significativos, como a propagação de informações falsas e a polarização de opiniões. 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]. O Twitter (atualmente X) 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, consequentemente, uma ótima fonte de dados, permitindo a extração de opiniões de grande parte da população.

Este estudo analisa a disseminação de informações nas redes sociais, onde opiniões são amplamente difundidas e alcançam diversos públicos. Os dados foram coletados do Twitter durante a Copa do Mundo de 2022 no Catar, um evento de grande importância para os brasileiros, unindo o país em torno do futebol. Foram analisados quase 12 milhões de tweets para entender as discussões, opiniões e perspectivas dos brasileiros sobre o evento. O estudo foca no envolvimento dos usuários e nas características dos

conteúdos compartilhados, explorando como o Twitter é utilizado para expressar opiniões e participar de discussões sobre temas relevantes da Copa do Mundo.

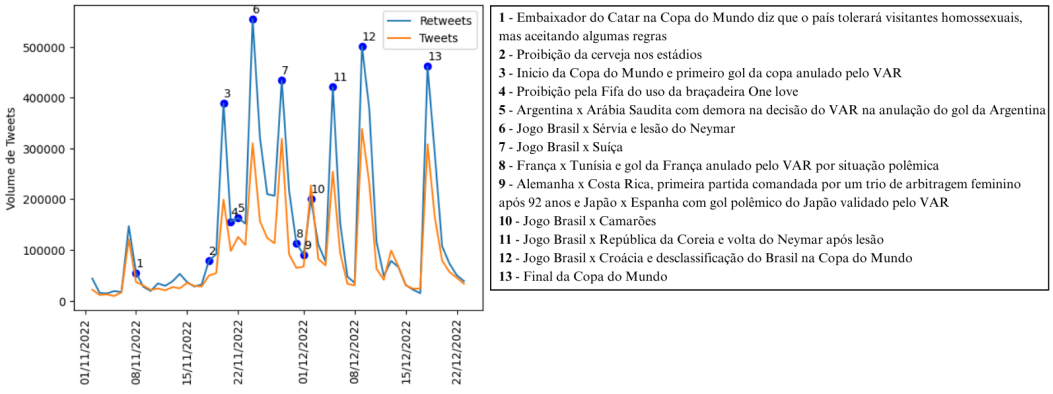
2. Trabalhos Relacionados

Os trabalhos descritos aqui abordam caracterização de dados em redes sociais, mais especificamente no Twitter. Técnicas de análise de sentimentos foram empregadas em [Teixeira and Azevedo 2011], para avaliar se informações coletadas do Facebook e Twitter que poderiam ser utilizadas para prever valores comerciais de produtos ou serviços antes de lançamento no mercado. As empresas reconheceram o valor dessas plataformas para promover seus produtos e analisar a percepção do público em relação a eles. Em [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 do debate sobre a vacinação, destacando como as pessoas utilizam o Twitter para compartilhar suas impressões e preocupações sobre o tema. Outro trabalho de análise de dados foi feito em [Lins 2020], que 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.

3. Conjunto de Dados e Análises

O conjunto de dados utilizado é parte do Projeto <*haa*>[1]] e contempla *tweets* que mencionam uma das seguintes 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, abrangendo o intervalo de 01.11 a 31.12 de 2022.

Figura 1. Volume de *tweets* durante o período coletado



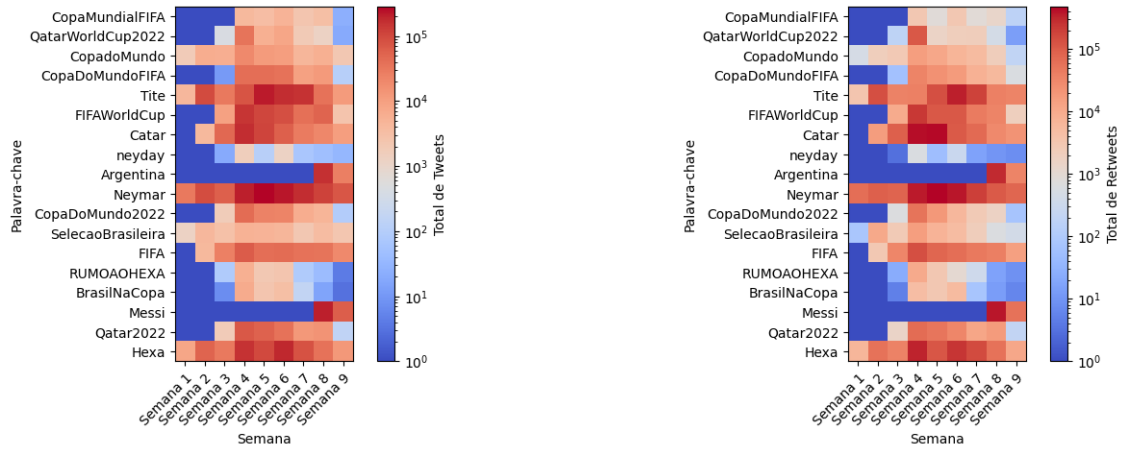
3.1. Análise temporal e palavras populares

A Figura 1 apresenta uma análise da evolução temporal das discussões relacionadas à Copa do Mundo de 2022. Através da série temporal dos números diários de *tweets* e

[1] *hidden for anonymous authorship*

retweets, é possível observar a ocorrência de picos significativos que coincidem com eventos relevantes durante o período, tal como a proibição da cerveja nos estádios. O primeiro pico significativo ocorreu no primeiro dia da Copa do Mundo, durante a cerimônia de abertura e início dos jogos. Os picos numerados como 6, 7, 10, 11 e 12 correspondem aos dias em que o Brasil jogou. O pico 6 foi o maior deles, pois marcou a estreia da seleção brasileira e coincidiu com a lesão de Neymar. O pico 10 ocorreu em um jogo em que o Brasil já estava classificado para as oitavas de final. Já o pico 12 foi o segundo maior e ocorreu no jogo em que o Brasil foi desclassificado. O pico 13 correspondeu à final da competição e foi o terceiro maior pico observado. Os picos no gráfico estão presentes nas linhas de *tweets* e *retweets*, ocorrendo sempre em conjunto. Isso indica como os eventos impulsionam o debate durante o período analisado.

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



(a) Tweets (b) Retweets

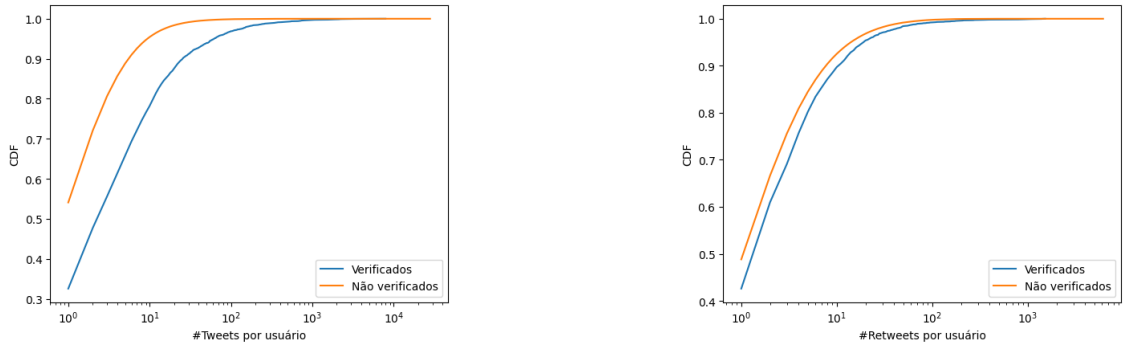
A popularidade de uma palavra-chave em uma dada semana é proporcional ao número total de *tweets* ou *retweets* que a mencionam. A Figura 2) mostra esta análise através de mapas de calor, onde cada célula representa o total de *tweets* ou *retweets* que mencionam cada palavra-chave em uma dada semana, utilizando uma escala logarítmica. A palavra-chave *Neymar* foi a mais popular e mais alta da quarta a sexta semana, período em que a Seleção Brasileira jogou e quando o jogador sofreu lesão, se recuperou e voltou a jogar ainda durante a competição. Outra palavra-chave bastante citada nas semanas quatro e cinco foi *Catar*, sede do evento. *Tite* e *Hexa* também foram palavras bastante mencionadas. É interessante destacar que as palavras-chave *Messi* e *Argentina* não registraram nenhuma menção durante a maior parte do período, mas tiveram um aumento significativo de citações no final, durante a oitava semana, momento do torneio em que a Argentina, com *Messi* como jogador, saiu vitoriosa.

3.2. Perfil dos Indivíduos

Os indivíduos foram categorizados em contas verificadas, maior interesse público e participação intensa em conversas sobre eventos relevantes para a sociedade, e não verificadas. A Figura 3 ilustra as distribuições de probabilidade acumulada dos números de *tweets* e *retweets*. Nos dados coletados as contas verificadas têm maior tendência a postar mais *tweets* e *retweets*. Cerca de 90% dos indivíduos com contas verificadas publicam até 30 *tweets* e 12 *retweets*, enquanto 90% 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 3. Distribuições dos números de tweets e retweets por tipo de conta

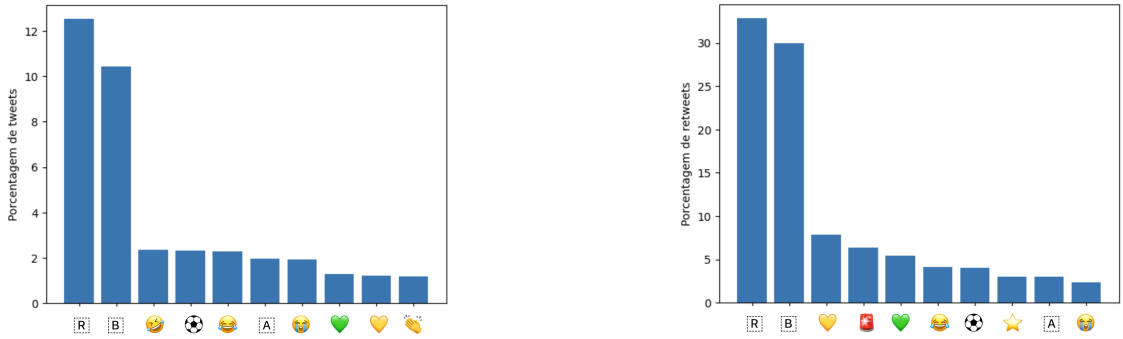


(a) Tweets (b) Retweets

3.3. Análise dos Emojis

Os *emojis* são usados nas redes sociais para complementar a comunicação, e representam emoções, ideias ou simbolismos, e 16,8% dos *tweets* e 33,6% dos *retweets* continha pelo menos um *emoji*.A Figura 4 apresenta os *emojis* mais utilizados, sendo o mais popular o do desenho da letra *R*, amplamente utilizado em *tweets* relacionados ao Brasil e à Argentina, representando as abreviações *BR* e *AR*. Os *emojis* de desenho das letras *B* e *A* também foram 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 e a bola de futebol, o que era esperado dado o contexto da competição.

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



(a) Tweets (b) Retweets

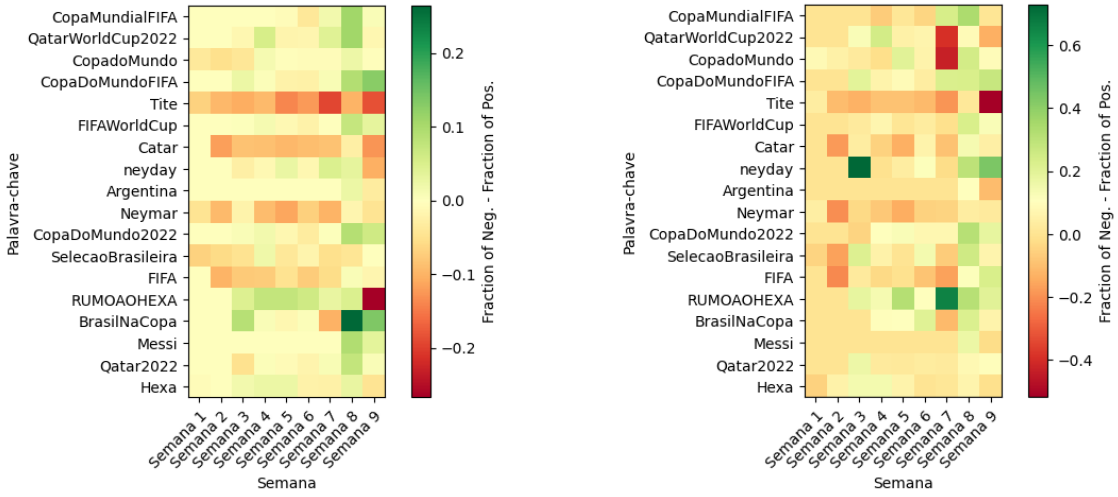
A frequência elevada dos *emojis* de choro pode expressar tanto tristeza quanto felicidade. O choro de tristeza foi utilizado em decorrência da derrota do Brasil no torneio e das lesões sofridas por alguns jogadores brasileiros, que resultaram em desfalques na equipe. Por outro lado, os *emojis* de choro 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. A análise de *emojis* indica que os símbolos 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.

3.4. Análise de Sentimentos

Consideramos *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. A fim de entender como o sentimento expresso pelos usuários varia em relação a cada palavra-chave, contrastamos *tweets* e *retweets* classificados como positivos e negativos que mencionam as palavras-chave ao longo das semanas.

Figura 5. Evolução semanal da diferença entre a fração de sentimentos positivos e negativos



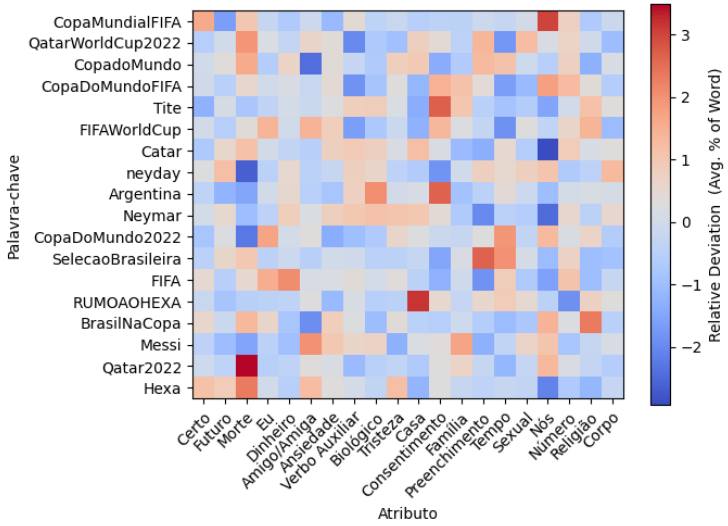
(a) Tweets (b) Retweets

O mapa de calor apresentado na Figura 5 resume essas diferenças através da pontuação contrastiva, sendo calculada como a diferença entre a fração de *tweets* positivos e negativos. Os *posts* que mencionam a palavra-chave *Tite* tendem a ser mais negativos, principalmente nas últimas semanas. 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, tendo relação com a desclassificação do Brasil. Outras duas palavras-chave com fração ligeiramente negativa foram *Catar* e *Neymar*, o que está provavelmente ligado a críticas. Os *tweets* com maior fração de sentimento positivo tendem a ser mais para as últimas semanas, observado em palavras-chave mais genéricas *CopaMundialFIFA*, *QatarWorldCup2022*, *CopaDoMundoFIFA*, *FIFAWorldCup*, *CopaDoMundo2022*, *BrasilNaCopa*, *Messi* e *Qatar2022*. Foi analisada também a pontuação contrastiva dos *retweets*, que podem ser considerados como uma medida da intensidade de difusão da informação nas redes sociais. A Figura 5(b) mostra que, em geral, os *retweets* seguem um padrão semelhante aos *tweets*. No entanto, na semana sete, 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 *RUMOAOHEXA* apresentou um sentimento mais positivo nos *retweets* em comparação aos *tweets*. 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.

3.5. Análise Psicolinguística

As análises psicolinguísticas foram feitas através da categorização das palavras em diferentes atributos relacionados ao estilo linguístico, conceitos afetivos e cognitivos, amplamente utilizado para esse tipo de análise. A frequência média desses atributos foi calculada para cada palavra-chave nos *tweets* e *retweets*, nos quais foram identificados 64 atributos disponíveis. Em seguida, foram analisadas diferenças estatísticas entre os debates em torno de diferentes palavras-chave, explorando a frequência média dos atributos associados a cada palavra-chave. O teste não paramétrico de Kruskal-Wallis [Kruskal and Wallis 1952] foi usado para selecionar atributos que apresentassem diferenças significativas entre as palavras-chave. Para lidar com a grande quantidade de atributos, o coeficiente de Gini [Yitzhaki 1979] foi usado para selecionar, dentre os 64, os 20 mais discriminantes. Os atributos mais relevantes foram identificados para cada palavra-chave através de um mapa de calor, considerando todos os *tweets* e *retweets*. A Figura 6 mostra um mapa de calor com essa análise.

Figura 6. Top LIWC atributos extraídos dos *tweets* e *retweets* coletados



Os resultados mostram diferenças nos atributos selecionados com postagens relacionadas às palavras-chave. Por exemplo, as postagens associadas a termos amplos, como *CopaMundialFIFA*, *QatarWorldCup2022*, *CopaDoMundo* e *Qatar2022*, frequentemente usam palavras relacionadas à morte. Além disso, *CopaMundialFIFA* e *CopaDoMundoFIFA* têm postagens que utilizam com frequência palavras relacionadas à coletividade. 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 empregam palavras relacionadas à ansiedade, coletividade, morte, religião, amizade, certeza e, principalmente, a palavra *casa* no caso de *RUMOAOHEXA*.

4. Conclusões e Trabalhos Futuros

Este trabalho explorou análises de dados da Copa do Mundo de 2022, incluindo perfis individuais, uso de emojis, sentimentos e psicolinguística. Utilizamos ferramentas como LeIA, LIWC, teste de Kruskal-Wallis e coeficiente de Gini para obter diferentes perspectivas sobre as discussões. As análises revelaram como eventos externos influenciam as

interações online, formando redes de opiniões alinhadas. Isso destaca a internet como reflexo das relações sociais reais e a importância do estudo na compreensão do impacto das dinâmicas online na formação de opinião sobre eventos globais. Como trabalhos futuros, são previstas: (i) geração de uma rede com conjunto de dados completo; (ii) extração de *backbone* da rede com todo o conjunto de dados; (iii) detecção de comunidades do conjunto de dados completo; comparação de comunidades detectadas com diferentes algoritmos; análise das comunidades detectadas.

Referências

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.

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.

Teixeira, D. and Azevedo, I. (2011). Análise de opiniões expressas nas redes sociais. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (8).

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.