Tweet_Eleições_2022: Um *dataset* de *tweets* durante as eleições presidenciais brasileiras de 2022

Luciano José da Silva[1], Livia A. dos Santos[2], Renata Araujo[1,2,3], Orlando B. Coelho *(in memoriam)*[2], **Ana Grasielle D. Correa[1,2], Ivan Carlos A. Oliveira[2]**

[1]Programa de Pós-Graduação em Computação Aplicada - Universidade Presbiteriana Mackenzie - São Paulo – SP – Brasil

[2]Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie - São Paulo - SP - Brasil.

[3]Programa de Pós-Graduação em Sistemas de Informação - EACH/USP - São Paulo - SP - Brasil

luciano.jose-silva@outlook.com, liviaalabarse.santos@mackenzista.com.br, {renata.araujo, orlando.coelho, ana.correa, ivan.oliveira}@mackenzie.br

Abstract. *In this paper we present Tweet_Eleições_2022, a dataset containing tweets related to political episodes that occurred during the 2022 Brazilian presidential elections. The dataset is mainly applicable to research that is interested in data from this Brazilian historical period, and has significant value for research that uses social network data in Brazilian Portuguese.*

Resumo. *Neste artigo apresentamos o dataset Tweet_Eleições_2022, contendo tweets relacionados a episódios de cunho político ocorridos durante as eleições presidenciais brasileiras de 2022. O dataset é aplicável principalmente a pesquisas que tenham o interesse em dados desse período histórico brasileiro, e possui valor significativo para pesquisas que façam uso de dados de redes sociais em português brasileiro.*

1. Introdução

As disputas eleitorais têm sido impactadas diretamente pelo fenômeno das redes sociais. Esses ambientes se tornaram férteis para observar fenômenos como a ultra polarização das discussões políticas, a radicalização e a manipulação de fatos dentro de "bolhas" de influência [Fisher 2023]. Diversos estudos na área de análise de redes sociais se debruçam sobre o desafio de acompanhar o debate político que se estabelece nas plataformas, principalmente aquelas com maior representatividade e número de usuários [Recuero et. al. 2019][Hong e Nadler 2012].

O comportamento das redes sociais nessas plataformas durante períodos eleitorais tem sido objeto de interesse de pesquisas no cenário brasileiro [De Paula Filho e Garcia 2015][Martins et. al. 2019][Nobre et. al. 2019][Reis e Benevenuto 2022]. O ano de 2022 é considerado um dos períodos eleitorais mais polarizados da nossa história, com grande participação das de redes sociais no cenário do debate público, servindo como base para estudos [Oliveira e Oliveira 2023][Paiva et.al. 2023][Pereira et. al. 2023][Santana et. al. 2023][Kappaun e Oliveira 2023][Santos e Berton 2023][Pinto e Silva 2023][Braga et. al. 2022][Gadelha et. al. 2023][Silva e Faria 2023].

Dentro desse contexto, a pesquisa desenvolvida em Silva et. al. (2023) teve como objetivo monitorar dados do *Twitter*[1] ao longo do processo eleitoral para presidência do Brasil em 2022. Foi desenvolvido um *pipeline* de análise de *tweets* para evidenciar a dinâmica de comportamento das redes sociais de acordo com a ocorrência de eventos politicamente relevantes no cenário nacional e noticiados na imprensa. Ao longo do ano de 2022, foram extraídos cerca de 26Gb de dados de *tweets* relacionados a diversos eventos ocorridos durante a campanha eleitoral. Os dados coletados nesta pesquisa foram organizados no *dataset Tweet_Eleições_2022*.

2. Trabalhos Relacionados

Apesar das diversas pesquisas relatando fenômenos nas redes sociais no Brasil, durante as eleições de 2022, o número de *datasets* disponibilizados derivados dessas pesquisas são em número menor. Paiva et. al. (2022), coletaram e disponibilizaram 780 mil *tweets* em português no período de set/2022 a fev/2023 para o estudo do debate sobre o feminismo durante as eleições. Pinto e Silva (2023) estudaram a caracterização de grupos políticos durante as eleições no intervalo entre 16/10 a 6/11/2022, a partir da coleta de 12.416 mensagens no Telegram. Segundo os autores, os dados serão disponibilizados no site do projeto *Online Political Polarization*, onde outros conjuntos de dados sobre eleições estão também disponibilizados. Braga et. al. (2022) estudaram os discursos dos perfis de mídia e de perfis comuns com base em 27.000 *tweets* coletados entre 04 e 15/07/2022, mas o conjunto de dados utilizados na pesquisa não foi disponibilizado. O diferencial do *dataset Tweet_Eleições_2022* está em seu volume (9.474.155 *tweets*) e em sua cobertura, contendo *tweets* de diferentes momentos ao longo de 2022, relacionados a acontecimentos veiculados na mídia oficial.

3. Criação do *Dataset*

O *dataset* descrito neste artigo foi criado a partir da API da plataforma *Twitter* durante o período das eleições de 2022 no Brasil, com o intuito de capturar *tweets* relacionados às eleições e aos temas políticos relevantes. A figura 1 ilustra o processo de construção do conjunto de dados, delineando os passos desde a coleta inicial até o *dataset* final.

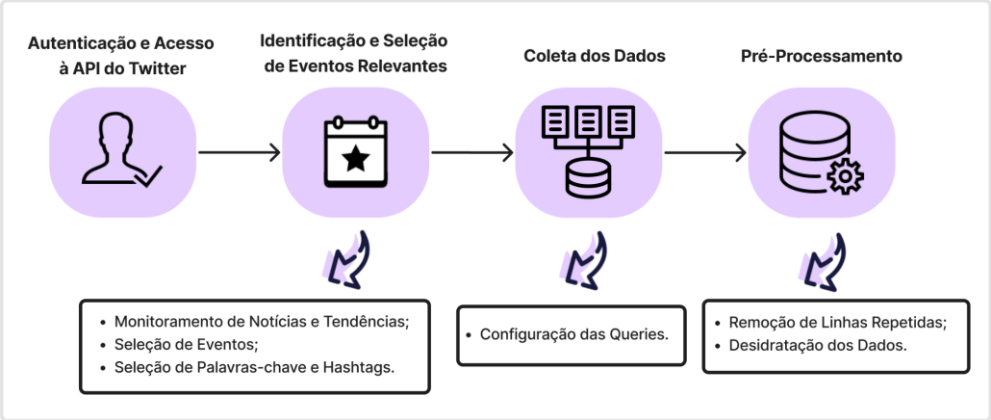

Figura 1. Fluxo de construção do *dataset Tweet_Eleições_2022*

Para a **coleta dos dados**, foi adotada uma abordagem que envolveu a configuração de um *token* de acesso fornecido pelo *Twitter* que é essencial para autenticar e autorizar

[1] Atualmente denominado *X*.

o acesso à API, permitindo assim a extração dos *tweets*. Com a autenticação garantida, o próximo passo foi identificar e selecionar as notícias e acontecimentos relacionados às eleições de 2022. Isso envolveu monitorar diversos canais de informação, como sites de notícias, portais on-line e jornais, a fim de identificar temas relevantes e *hashtags* em destaque. A seleção das notícias foi realizada com base no interesse e avaliação dos autores sobre a sua repercussão na mídia oficial e o quanto se associava direta ou indiretamente a personalidades ou a ideologias político-partidárias presentes nas eleições. A Tabela 1 mostra os eventos a partir dos quais os *tweets* foram extraídos e que compõem o *dataset*.

Tabela 1. Eventos relacionados aos *tweets* extraídos

Mês de coleta	Assunto
Maio 2022 (279.149 *Tweets*)	Bate-boca entre Ciro Gomes e Gregório Duvivier; Encontro de Bolsonaro e Elon Musk em SP; Atraso nas vacinas contra COVID-19 para crianças; Empresa ligada a shows de Gusttavo Lima e BNDES
Junho 2022 (2.008.769 *Tweets*)	Pesquisas Datafolha (Lula x Bolsonaro e falta de comida); Tensão entre Bolsonaro e STF sobre possibilidade de descumprimento de decisões; Ministro da Defesa pede reunião exclusiva com TSE; Tarcísio defende Bolsonaro em investigação sobre ex-ministro da educação; STF pressiona Nunes Marques por decisão sobre aliado bolsonarista; Barroso move ação contra Magno Malta por calúnia; Forças Armadas se sentem excluídas no debate sobre segurança das urnas; Desaparecimento de Dom Philips e Bruno Pereira; Temer alerta contra mudanças na Lei das Estatais; Ciro Gomes no Flow Podcast; Casagrande critica governo Bolsonaro; Bloomberg informa que Bolsonaro pediu ajuda a Biden para reeleição; Biden e Bolsonaro discutem fake news e eleições; Juíza impede aborto de menina de 11 anos estuprada; Aumento da gasolina e diesel; Renúncia do presidente da Petrobras; Monark defende o direito de existência de um partido nazista; Prisão do ex-ministro Milton Ribeiro; Mark Ruffalo critica Bolsonaro em relação a Amazônia; Assédio sexual na caixa Econômica; Aproximação do presidente do Bradesco com militares; Bens de Paola Silveira são bloqueados; Tiago Leifert opina sobre dilema entre Lula e Bolsonaro
Julho 2022 (439.512 *Tweets*)	Acusações de incitação à violência contra Bolsonaro; Pesquisa eleitoral BTG e Datafolha; Ataque ao Capitólio; Lula busca apoio de eleitores de centro; Dilma chama Temer de "golpista"; Monark defende Leo Lins; Renúncia de vice-presidente da Caixa por denúncias de assédio; Assassinato de guarda municipal petista por apoiador de Bolsonaro; Oficialização da candidatura de Simone Tebet à Presidência e opinião sobre direitos e segurança das mulheres; Reação exaltada de Carla Zambelli sobre fundo eleitoral.
Agosto 2022 (1.605.485 *Tweets*)	Alexandre de Moraes se torna presidente do TSE; Troca de mensagens entre Augusto Aras e empresários bolsonaristas; Carta em defesa da democracia é lida em evento da USP; Ciro Gomes é entrevistado no Roda Viva e Jornal Nacional; Debate entre os candidatos ao governo de São Paulo e do Rio de Janeiro; Empresários bolsonaristas são investigados por estimular um suposto golpe de Estado; Bolsonaro ataca jornalista Vera Magalhães

Setembro 2022 (1.412.042 *Tweets*)	Casagrande critica Neymar por votar em Bolsonaro; Ciro gomes critica campanha pelo voto útil; Entrevista de Lula para CNN; Pesquisa eleitoral Datafolha, Ipec e BTG; Debate presidencial da Globo e SBT; Deputado Douglas Garcia ofende Vera Magalhães em debate de candidatos ao governo de SP; Brasileiro tenta assassinar Cristina Kirchner; Paraná pesquisas recebe 2,7 m de partido de Bolsonaro; Teto de intenções de voto de Bolsonaro é analisado; Morte da Rainha Elizabeth II
Outubro 2022 (2.014.753 *Tweets*)	Debate sobre segurança e confiabilidade das urnas; Início da votação nas Eleições 2022; Roberto Jefferson lança granada contra policiais federais e é preso
Janeiro 2023 (1.714.445 *Tweets*)	Tentativa de golpe no dia 08/01; Declaração de emergência em saúde pública no território yanomami

Na sequência, foram selecionadas palavras-chave e *hashtags* pertinentes aos eventos escolhidos. Esses termos foram utilizados para configurar a *query* de extração, especificando o período de interesse, o limite máximo de *tweets* a serem recuperados, e diversos campos do *tweet* foram coletados para fins da pesquisa para qual o *dataset* foi inicialmente gerado [Silva et. al., 2024]. O *dataset* gerado após o pré-processamento conta com quatro colunas (Tabela 2).

Foram realizadas duas etapas de **pré-processamento** para garantir a qualidade, consistência e requisitos de privacidade dos dados: **1) Remoção de Linhas Repetidas**: remoção de linhas repetidas originadas pela coluna *mention_username*. Para cada usuário citado no *tweet*, uma nova linha era criada com o resto das colunas repetidas. Para lidar com esse problema, foi implementado um processo de agrupamento de *tweets* com base na coluna *conversation_id*; **2) Desidratação dos Dados:** Foram retiradas as identificações diretas de perfis e os conteúdos dos *tweets*, preservando a privacidade dos usuários. A desidratação evita qualquer potencial de re-identificação sem o uso da *API* para a reidratação dos dados. Considerando o intervalo de tempo entre 27 de Abril de 2022 e 23 de Janeiro de 2023, o resultado foi um *dataset* composto por cerca de 9.5 milhões de *tweets*, representando uma variedade de perspectivas e discussões relacionadas às eleições de 2022 no Brasil. O número de *tweets* coletados a cada mês pode ser visualizado na Figura 2. Os dados coletados foram **armazenados** em repositórios com livre acesso[2].

4. Aplicações

O *Tweet_Eleições_2022* é aplicável principalmente a pesquisas que tenham o interesse de analisar dados deste período da história brasileira, mas também possui enorme valor para qualquer pesquisa que faça uso de dados de redes sociais em português brasileiro. O *dataset* foi criado e utilizado em projeto de pesquisa voltado ao acompanhamento da rede social *Twitter* e identificação de principais atores e dinâmica de polarização na rede durante as eleições de 2022 [Silva et. al 2024]. Outro uso para o *dataset* será no contexto de projeto de pesquisa voltado à aplicação de mineração de argumentos para identificação

[2] https://zenodo.org/records/11206577

e visualização de discussões em redes sociais.

Campo	Tipo	Descrição
Created_at_convert	*date (ISO 8601)*	Data da criação do *tweet*
author_id	*string*	Identificador único do autor do *tweet*
conversation_id	*string*	ID do *tweet* que iniciou a conversa
referenced_tweets	*string*	ID do *tweet* mencionado. Se o *tweet* for original, esse último campo terá valor nulo

Tabela 2. Dicionário de dados do *dataset* Tweet_Eleições_2022

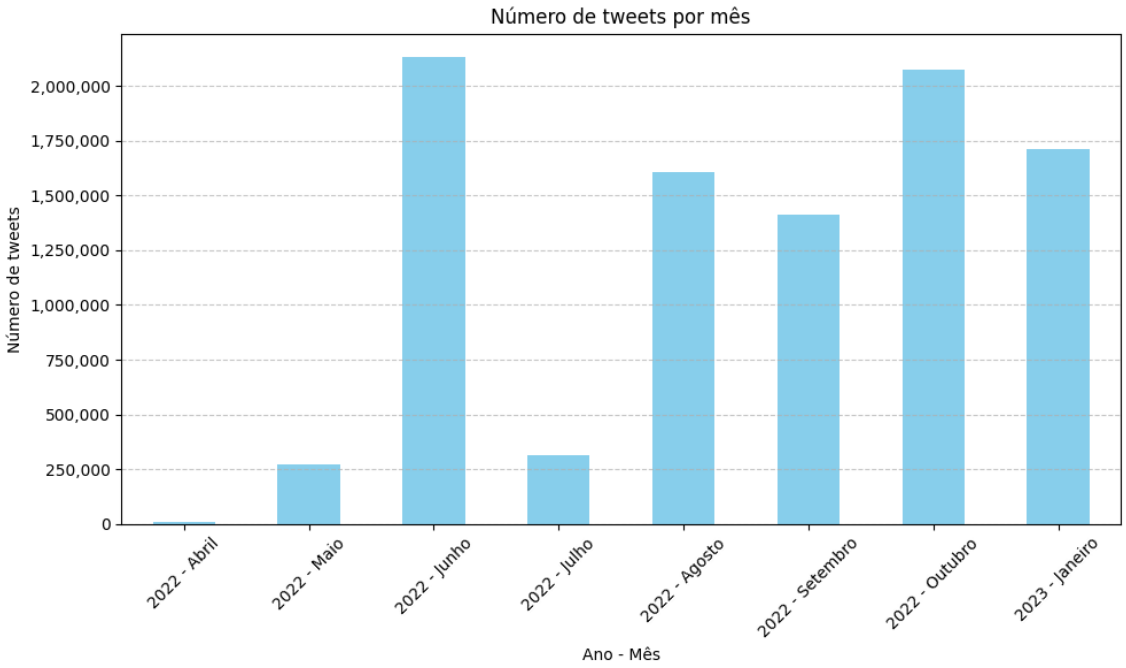

Figura 2. Gráfico de número de *tweets* por período (mês) contidos no *dataset*

5. Considerações Finais

As pesquisas na área de análise de redes sociais utilizando dados brasileiros e em língua portuguesa requerem o acesso aos dados das plataformas e a disponibilidade de conjuntos de dados organizados para as análises. Com a interrupção do acesso aberto a pesquisadores pelo *Twitter* no primeiro semestre de 2023, muitas pesquisas se viram prejudicadas no acesso a esses dados. Torna-se importante que a comunidade de pesquisa compartilhe os *datasets* produzidos em seus projetos.

A principal limitação do *dataset* decorre das questões éticas da disponibilização de dados de redes sociais. Segundo as regras de uso do *Twitter*, o uso dos dados obtidos

via API precisam garantir a privacidade de seus usuários, exigindo a retirada de informações que possam identificá-los, um processo conhecido como "desidratação". Desta forma, se a identificação dos usuários for algo necessário na pesquisa, o pesquisador precisará ter acesso à API para poder (re)hidratá-los.

Agradecimentos

Os autores agradecem o apoio financeiro da FAPESP, processos #2021/14772-1 e #2023/04042-1, e do CNPq processo #305645/2022-6.

References

Braga, F. T.; Dos Santos, I. M. e Mota, M. P. (2022) Uma Análise Comparativa sobre o que Dizem a Grande Mídia e os Usuários Comuns no Twitter sobre os Presidenciáveis Brasileiros em 2022. En: Workshop sobre Aspectos da Interação Humano-Computador na Web Social. Diamantina. Sociedade Brasileira de Computação. p. 79-86.

De Paula Filho, W. e Garcia, A. C. (2015). Predição do resultado das eleições presidenciais do Brasil baseado em tuítes. En: *Brazilian Workshop On Social Network Analysis And Mining*. Recife. Sociedade Brasileira de Computação.

Fisher, M. (2023). "A máquina do caos: como as redes sociais reprogramaram nossa mente e nosso mundo." São Paulo: Todavia.

Gadelha, T., Monteiro, J. M., Machado, J.. Claudino, I., Santos, R., Galick, L. e Santos, C. (2023). Ativismo da extrema direita brasileira no WhatsApp: O que mudou das eleições de 2018 para 2022?. En: Simpósio Brasileiro de Banco de Dados. Belo Horizonte/MG. Sociedade Brasileira de Computação. p. 336-341.

Hong, S. e Nadler, D. (2012). Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*. Elsevier, v. 29, n. 4, p. 455–461.

Kappaun, A. e Oliveira, J. (2023). Análise sobre Viés de Gênero no Youtube: Um Estudo sobre as Eleições Presidenciais de 2018 e 2022. Em: *Brazilian Workshop on Social Network Analysis and Mining*. João Pessoa/PB. Sociedade Brasileira de Computação. p. 127-138.

Martins, E. A., Gonçalves, K. C. e Miranda Filho, R. (2019). Caracterizando a campanha presidencial brasileira em 2018 usando dados do Twitter. En: *Brazilian Workshop on Social Network Analysis and Mining,* 2019, Belém. Sociedade Brasileira de Computação. p. 131-142.

Nobre, G. P., Almeida, J, M. e Ferreira, C. H. G. (2019). Caracterização de bots no Twitter durante as Eleições Presidenciais no Brasil em 2018. En: *Brazilian Workshop on Social Network Analysis and Mining*. Belém. Sociedade Brasileira de Computação. p. 107-118..

Oliveira, B. e Oliveira, L. S. (2023). "Aprendizado de Máquina e Análise de Sentimento em Redes Sociais: Um Estudo de Caso Usando as Eleições Presidenciais em 2022". Em: *Congresso Latino-Americano de Software Livre e Tecnologias Abertas*. Foz do Iguaçu/PR. Sociedade Brasileira de Computação, p. 107-112.

Paiva, B. F., Barbosa, B. R. G, Silva, A. P. C. e Moro, M. M. (2023). “O debate do feminismo no Twitter: Um estudo de caso das eleições brasileiras de 2022”. Em: *Brazilian Workshop on Social Network Analysis and Mining*. João Pessoa/PB. Sociedade Brasileira de Computação, p. 103-114.

Pereira, R., Alves, A., Vidal, D., Moura, F., Cabral, L., Paulino, R., Serrufo, M. e Figueiredo, K. (2023). Análise de Sentimento de Postagens de Usuários no Twitter Combinando GPT-3 e Aprendizado de Máquina: Um Estudo de Caso Sobre o 2º Turno das Eleições Presidenciais Brasileiras. En: *Workshop sobre Aspectos da Interação Humano-Computador na Web Social*. Maceió/AL. Sociedade Brasileira de Computação. p. 20-27.

Pinto, J. S. e Silva, T. H.. (2023). Caracterização de Grupos Políticos no Telegram Durante a Eleição Presidencial de 2022. En: Concurso de Trabalhos de Iniciação Científica - Simpósio Brasileiro De Sistemas Multimídia e Web. Ribeirão Preto/SP. Sociedade Brasileira de Computação. p. 55-58.

Recuero, R., Zago, G. e Soares, F. (2019). Using social network analysis and social capital to identify user roles on polarized political conversations on twitter. Social media+ society. SAGE Publications. London, England, v. 5, n. 2, p. 2056305119848745, 2019.

Reis, J. C. S. e Benevenuto, F. (2022). Detecção Automática de Desinformação em Diferentes Cenários: Eleições nos Estados Unidos e no Brasil. Em: *Brazilian Workshop on Social Network Analysis and Mining.* Niterói. Sociedade Brasileira de Computação. p. 1-12.

Santana, M., Lima, J., Correa, A. e Brito, K.. (2023). Engajamento no TikTok dos candidatos às eleições Brasileiras de 2022 – Resultados Iniciais. Em: *Brazilian Workshop On Social Network Analysis And Mining*. João Pessoa/PB. Sociedade Brasileira de Computação. p. 151-162.

Santos, D. K. S. e Berton, L. (2023). Analysis of Twitter users' sentiments about the first round 2022 presidential election in Brazil. In: *Encontro Nacional de Inteligência Artificial e Computacional.* Belo Horizonte/MG. Sociedade Brasileira de Computação p. 880-893.

Silva, L. J., Araujo, R. M., Correa, A. G. D. (2024). Pipeline para monitoramento de discussões políticas no Twitter: estudo de caso com o evento político de 8 de janeiro de 2023. Em: Brazilian Workshop On Social Network Analysis And Mining. Brasília/DF. Sociedade Brasileira de Computação.

Silva, S. M. B. e Faria, E. R. (2023). Análise de sentimentos expressos no Twitter em relação aos candidatos da eleição presidencial de 2022. Em: *Brazilian Workshop on Social Network Analysis and Mining*. João Pessoa/PB. Sociedade Brasileira de Computação. p. 79-90.