Caracterização e Predição de Usuários Tóxicos no Twitter/X durante as Eleições Brasileiras de 2022

Samuel Lopes Pinto[1], **José Julio Campolina**[1], **João Pedro M. Sena**[1], **Gabriel Félix**[1], **Lucas N. Ferreira**[1], **Julio C. S. Reis**[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV) – Brasil

{samuel.r.pinto, josecampolina, joao.sena}@ufv.br

{gabriel.felix, lucas.n.ferreira, jreis}@ufv.br

Abstract. *With the emergence of smartphones, social platforms have become widely popular due to their ease of use. These platforms provide a conducive environment for communication between people on various topics. Especially in the political context, these platforms have been widely used to carry out virtual electoral campaigns and disseminate illicit content, including hate speech. In this context, computational solutions can be useful for early identification of this type of message. We explored posts from Twitter/X users to propose an approach that uses a pre-trained BERT model for Brazilian Portuguese (BERTimbau), to identify potentially toxic users considering the Brazilian political context. Our best results highlight that it is possible to achieve around 85% in terms of F1 score in the task of identifying a potentially toxic users. Therefore, in addition to contributing to the understanding of the characteristics of toxic speech on Twitter/X, this study highlights the potential of machine learning approaches to identify users with inappropriate behavior in the online environment, which can be useful to mitigate the impact caused by propagation of this type of content in these environments.* ***Warning! This paper contains offensive words and tweet examples.***

Resumo. *Com o surgimento dos smartphones, as plataformas sociais tornaram-se amplamente populares devido à sua facilidade de uso. Essas plataformas fornecem um ambiente propício para a comunicação entre pessoas sobre diversos assuntos. Especialmente no contexto político, essas plataformas têm sido amplamente utilizadas para realização de campanhas eleitorais virtuais e disseminação de conteúdo ilícito, incluindo discurso de ódio. Neste contexto, soluções computacionais podem ser úteis para identificação precoce deste tipo de mensagem. Exploramos publicações de usuários do Twitter/X para a proposição de uma abordagem que utiliza um modelo BERT pré-treinado para o português brasileiro (BERTimbau), para identificação de usuários potencialmente tóxicos considerando o contexto político brasileiro. Nossos melhores resultados obtiveram cerca de 85% em termos de F1 score na tarefa de identificar um usuário potencialmente tóxico. Logo, além de contribuir para a compreensão das características do discurso tóxico no Twitter/X, este estudo releva o potencial das abordagens de aprendizado de máquina para identificar usuários com comportamento inadequado no ambiente online, o que pode ser útil para mitigar o impacto causado pela propagação desse tipo de conteúdo nesses ambientes.* ***Aviso! Este artigo contém palavras e exemplos de tweets ofensivos.***

1. Introdução

Com o advento dos *smartphones*, plataformas de mídias sociais se tornaram sistemas amplamente populares devido ao baixo custo e à sua facilidade de uso. Essas plataformas se consolidam em um ambiente extremamente propício para a comunicação entre pessoas sobre diversos assuntos, incluindo esportes, saúde e política [Conover et al. 2011, Christhie et al. 2018, Reis et al. 2023a]. Um problema é que, especialmente no contexto político, esses ambientes têm sido amplamente explorados para disseminação de conteúdo ilícito [Guimaraes et al. 2022, Araujo et al. 2023], incluindo campanhas de desinformação [Reis et al. 2023b] e discurso de ódio [da Fonseca et al. 2024]. Todos esses fatores, além de promoverem a distorção do debate político no ambiente online, tem contribuído para um aumento significativo da polarização dos eleitores brasileiros. Nas eleições presidenciais de 2022, por exemplo, foram divulgados vários casos de violência política, com alguns deles, inclusive, culminando em mortes [Queiroga 2022].

Por um lado, essas plataformas oferecem aos usuários um mecanismo bastante útil para que eles possam se expressar acerca de determinado assunto e/ou ideia ou ainda, manifestar a sua aprovação ou desaprovação em relação a determinado candidato e/ou suas propostas no cenário político. Por outro, conforme mencionado anteriormente, esses ambientes criam um espaço propício para a postagem de mensagens pejorativas ou tóxicas que podem ter reflexos negativos em várias esferas da sociedade, impactando diretamente a democracia e a vida das pessoas. Diante desse cenário, é fundamental compreender as características do discurso potencialmente tóxico presente nessas plataformas de mídias sociais, a fim de que, no futuro, sejamos capazes de propor soluções computacionais que sejam úteis e efetivas na mitigação do impacto negativo ocasionado pela disseminação desse tipo de conteúdo nesses ambientes, eventualmente viabilizando a identificação e punição dos reais responsáveis. Especialmente no contexto brasileiro, estes mecanismos ainda são bastante escassos. É sobre esta lacuna que estabelece-se o objetivo deste estudo.

Particularmente, o objetivo deste trabalho é fornecer uma caracterização do conteúdo postado por usuários potencialmente tóxicos durante as eleições presidenciais brasileiras de 2022. Para isso, a partir de um conjunto de palavras-chave previamente definido, coletamos com o uso da API de dados históricos do Twitter/X, $\approx$ 6.9 milhões de *tweets* postados por cerca de 1.7 milhões de usuários únicos. Esses usuários foram ordenados com base na utilização de palavras que podem desencadear discussões tóxicas e o resultado desta etapa foi utilizado para separação deles em dois grupos distintos, que no contexto deste estudo são referenciados como "mais tóxicos" e "menos tóxicos". Em seguida, foi realizada uma coleta adicional do conteúdo postado pelos usuários destes dois grupos (i.e., *timeline*). A partir disso, exploramos uma representação vetorial do texto para captura do significado semântico e o contexto das palavras escritos por estes usuários no idioma português do Brasil (i.e, BERTimbau) e investigamos o potencial de 5 abordagens distintas baseadas em aprendizado de máquina na tarefa de identificar um usuário potencialmente tóxico. Por fim, também avaliamos, de forma preliminar, o potencial prático da abordagem proposta.

Em resumo, nossos resultados revelam características interessantes do conteúdo postado por usuários potencialmente tóxicos em plataformas de mídias socias. Além disso, apresentamos evidências de que as abordagens investigadas tem potencial interessante para identificação desses usuários no Twitter/X. O melhor modelo (i.e., XGB)

obteve $0,85$ e $0,94$ em termos de *F1 score*, e ROC AUC, respectivamente, evidenciando a utilidade deste tipo de mecanismo como ferramenta de auxílio para detecção desses usuários com comportamento eventualmente inadequado. Finalmente, nossa análise do potencial prático da estratégia proposta revelou que, ao utilizar dados do primeiro turno das eleições para treinamento, seria possível evitar a ocorrência de 103.677 publicações realizadas por um total de 313 usuários potencialmente tóxicos no período entre os turnos. Esperamos que nossas descobertas possam contribuir para a contenção do problema no ambiente online.

As seções subsequentes estão organizadas da seguinte forma. Na Seção 2, apresentamos trabalhos relacionados, enquanto na Seção 3 descrevemos o processo relativo a construção da base de dados explorada neste estudo. Uma caracterização dos dados coletados é apresentada na Seção 4. Depois, na Seção 5, descrevemos nossa abordagem proposta para identificação de usuários potencialmente tóxicos no Twitter/X e discutimos os principais resultados obtidos. Por fim, na Seção 6 concluímos o trabalho e apresentamos direções para pesquisas futuras neste contexto.

2. Trabalhos Relacionados

Pesquisas acerca do discurso de ódio em plataformas de mídias sociais têm sido objeto de diversas investigações nos últimos anos [Al-Hassan and Al-Dossari 2019, Lima et al. 2020, Teixeira and Reis 2023]. Parte desse crescente interesse deve-se ao fato de que a Internet, e consequentemente esses aplicativos, tornaram-se uma parte fundamental para o estabelecimento das relações sociais, transformando-se em uma importante ferramenta para entender a dinâmica das interações entre indivíduos e grupos em nossa sociedade, cada vez mais conectada. A análise das diversas ferramentas e das formas de dispersão do ódio online, como destacado por [Zannettou et al. 2018], é crucial para compreender e abordar efetivamente esse fenômeno crescente.

No contexto específico do Twitter/X, uma das principais plataformas de interação entre usuários em escala global, observa-se uma ampliação da pesquisa voltada para o entendimento, detecção e mitigação do discurso de ódio. Trabalhos notáveis, como os de [Davidson et al. 2017], destacam a complexidade desse desafio e ressaltam a importância de estratégias eficazes para combater as atitudes tóxicas presentes nesse ambiente virtual. Ao compreender as dinâmicas específicas do Twitter/X, os pesquisadores podem contribuir significativamente para o desenvolvimento de abordagens e políticas mais eficazes, promovendo um ambiente online mais saudável e inclusivo.

O avanço na detecção do discurso de ódio em plataformas de mídias sociais está sendo impulsionado por diversas abordagens inovadoras. Um exemplo destacável é o trabalho apresentado em [Almerekhi et al. 2020], que se destaca por identificar marcadores linguísticos específicos que desencadeiam toxicidade em discussões online. Essa pesquisa representa uma incursão valiosa no entendimento das sutilezas da linguagem digital, visando antecipar o risco futuro de publicação de discursos de ódio. Ao focar na modelagem em nível de usuário para identificar esses potenciais riscos relacionados a eventos específicos, o estudo contribui para preencher uma lacuna na pesquisa, destacando a importância de uma abordagem proativa na mitigação do ódio online.

Ainda neste contexto, [An et al. 2021] investigou o desenvolvimento de preditores do discurso de ódio contra asiáticos em plataformas sociais durante a pandemia de

COVID-19. O estudo empregou técnicas avançadas de processamento de linguagem natural para caracterizar usuários do Twitter/X que publicaram mensagens contendo tom xenofóbico contra asiáticos durante esse período. Ao comparar dois grupos de usuários - aqueles que postaram insultos anti-asiáticos e aqueles que não o fizeram - com base em características mensuradas antes da pandemia, os autores demonstraram a capacidade de prever quem eventualmente publicaria tais insultos. Além disso, é importante destacar o estudo de [Silva and Freitas 2022] por utilizar o modelo BERT para a língua portuguesa, denominado BERTimbau, com o objetivo de classificar discursos de ódio em bases de dados em português.

Em suma, nosso esforço é complementar aos esforços anteriores no sentido de apresentamos uma caracterização de mensagens oriundas de plataformas sociais, a saber, o Twitter/X durante um período político. No entanto, um diferencial presente em nosso trabalho, consiste na investigação de ferramentas automatizadas para controle e responsabilização de usuários que se utilizam do ambiente com suposto alto nível de anonimato para fazerem uso indevido de seu direto à liberdade de expressão e fomentar um ambiente político hostil e pouco produtivo para o debate. A metodologia empregada, embora com inspirações no estudo apresentado em [An et al. 2021], traz como diferencial uma aplicação considerando o idioma português do Brasil, onde esforços ainda são escassos. Esperamos que ajudando a identificar usuários com comportamento inadequado neste cenário, seja possível mitigar o impacto ocasionado pela disseminação de conteúdos danosos em plataformas de mídias sociais.

3. Base de Dados

Nesta seção apresentamos a estratégia adotada para construção da base de dados explorada nesse trabalho.

Definição de Palavras-Chave. Primeiramente, foi necessário definir os termos (ou palavras-chave) para a construção da base de dados explorada neste estudo. Neste contexto, é importante mencionar que esses termos foram definidos com base em publicações realizadas por veículos de imprensa, os quais descrevem essas palavras como potencialmente tóxicas no contexto político brasileiro [Kertzman 2020, Neves 2020]. Além disso, é importante destacar que foram selecionados termos representativos dos dois espectros políticos mais expressivos (i.e., esquerda e direita). Para manter um equilíbrio, foram selecionados 5 (cinco) termos mais comuns para cada um dos grupos, sendo eles: “esquerdista”, “comunista”, “petralha”, “esquerdopata”, “mortadela” para grupos de esquerda e “gado”, “fascista”, “genocida”, “bolsominion”, “coxinha” para grupos de direita.

Coleta e Rotulação dos Usuários. Após a definição dos termos a serem utilizados na coleta, a API de dados históricos do Twitter/X[1] foi explorada para obtenção das publicações realizadas na plataforma antes e durante os 2 (dois) turnos das eleições presidenciais brasileiras de 2022, a saber: de 02/01/2022 a 02/10/2022 (período pré-eleitoral até o primeiro turno), e 03/10/2022 a 30/10/2022 (período entre turnos). No contexto deste estudo essas bases de dados, nomeadas como “Primeiro turno” e “Segundo turno”, respectivamente, foram construídas e exploradas para propósitos distintos. Embora ambas as bases tenham sido exploradas durante o processo de caracterização dos dados coletados, a base de dados “Primeiro turno” é o *lócus* da análise. Ela foi utilizada como base para construção

[1] `https://developer.twitter.com/en/docs/twitter-api`

Tabela 1. Sumário da base de dados explorada neste estudo.

Período	02/01 a 02/10	03/10 a 30/10	Total
#Tweets	5.044.236	1.863.582	6.907.818
#Usuários únicos	1.465.523	847.353	1.708.342

da abordagem proposta, enquanto a base de dados do "Segundo turno" foi explorada para investigação do potencial prático da estratégia. No total, coletamos cerca de 6,9 milhões de *tweets* realizados por aproximadamente 1,7 milhão de usuários distintos, conforme apresentado na Tabela 1. É importante mencionar que existe uma interseção de 604.534 entre as bases de dados de ambos os períodos, ou seja, primeiro e segundo turnos. A Figura 1 apresenta o volume de publicações coletadas considerando ambas as bases de dados, por palavra-chave. É possível notar que "gado" é o termo mais expressivo, com 131.295 ocorrências, seguido de "bolsominion" e "comunista", com 82.023 e 71.281, respectivamente.

Na Tabela 2 são apresentados exemplos de *tweets* depreciativos para cada uma das palavras-chave utilizadas durante a coleta. É possível observar citações à usuários com ofensas diretas e também a figuras políticas do período de coleta como a usuária '@SorayaThronicke', senadora, e '@frota', deputado. É importante mencionar que foram retiradas as menções a perfis que não pertencem a pessoas públicas.

Obtidos os *tweets*, os usuários foram ordenados de acordo com a frequência das palavras tóxicas (palavras-chave) utilizadas em suas publicações. A Figura 2 apresenta a função de distribuição acumulada (CDF) do número de palavras tóxicas contidas nas mensagens coletadas por usuário. Note que estão representados usuários que postaram *tweets* contendo ao menos uma palavra tóxica. Ademais, podemos observar que a maioria dos usuários ($\approx 80\%$) utilizou até 150 palavras tóxicas em suas postagens. Por outro lado, uma pequena fração ($\approx 5\%$) desses usuários incluíram mais que 800 palavras tóxicas em seus *tweets*.

Depois de ordenar os usuários com base na frequência de palavras tóxicas utilizadas em suas postagens, separamos os usuários em dois grupos distintos: "mais tóxicos", que compreendem os 5% dos usuários com as maiores frequências e os "menos tóxicos", que são compostos pelos 5% dos usuários com as menores. Assim, foram obtidos conjun-

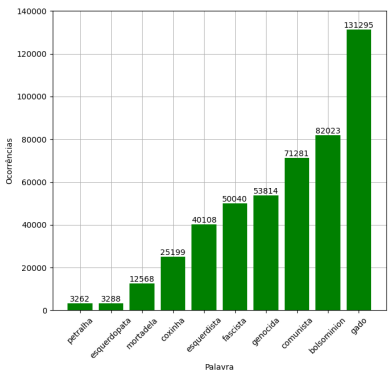

Figura 1. Frequência de palavras-chave na base de dados.

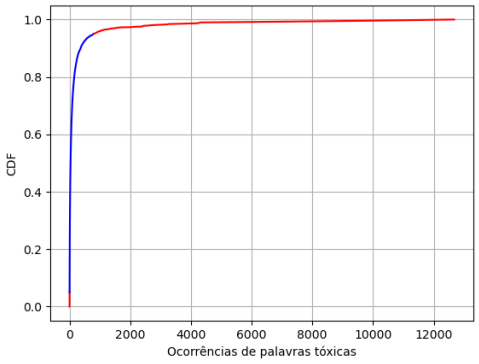

Figura 2. Função de distribuição acumulada (CDF) de ocorrências de palavras tóxicas por usuário.

Tabela 2. Exemplos de *tweets* tóxicos das palavras-chave avaliadas.

Palavra-chave	Exemplo
mortadela	"@user-1 @user-2 A mortadela tá enterrada no seu c* Pode olhar aí"
petralha	"É óbvio que o brocha do meu ex ia virar Cirista e ficar falando idiotice de petralha e não sei o que Por essas e outras que terminei com você seu otário"
esquerdista	"@user-1 @SorayaThronicke sei sim você chama esquerdista de burro e escreve errado por isso falei hípocrita Burro"
esquerdopata	"@user-1 @user-2 Saí do meu Brasil esquerdopata nordestino maldito"
comunista	"Sr Deus Por favor ilumina a cabeça de cada brasileiro amanhã de modo a não permitir que prevaleça a desgraça comunista corrupta abortista pródrogas próideologia de gênero desarmamentista totalitarista inimiga da liberdade do agronegócio [...]"
coxinha	"@user-1 Aquele coxinha neoliberal só vê à sua frente privatização ajuda aos poderosos prejuízo à classe trabalhadora e exaltação às empresas privadas Deus nos livre de gente assim"
bolsominion	"Cara algum bolsominion acha que vai virar voto fazendo carreata sábado de manhã Eu quero dormir seus filha da puta Que vontade de tacar ovo aqui de cima"
fascista	"@user-1 @frota Só sw for cm de rôla enfiada em teu rabo Nordeste vai ser mais de pro Lula seu gado inútil"
gado	"@user-1 @user-2 Festa que o degenerado sequestrou pra transformar em palanque pro gado otário."
genocida	"eu acho extremamente assustador ver esse monte de crnte fazendo corrente de oração p bolsoaro ganhar tipo você tá orando p um genocida filho da puta governar seu país"

tos com o mesmo número de usuários do topo e da base da lista ordenada de frequência, nesta ordem. Especificamente 3.443 usuários em cada grupo. Após a filtragem das contas públicas[2], foram obtidos 367 (do total 3.443 usuários) e 687 (também do total de 3.443) usuários em cada grupo (i.e., "mais tóxicos" e "menos tóxicos"), com 538.286 e 578.673 *tweets* nesta ordem, representados na Figura 2 pelas partes superior e inferior na cor vermelha, respectivamente. Em suma, considerando o período 02/01/2022 a 30/10/2022 (i.e., "Primeiro turno", usuários "mais tóxicos" postaram em média 1.480 *tweets*, enquanto os "menos tóxicos" fizeram 857 postagens com o uso de palavras tóxicas.

4. Caracterização da Base de Dados

Uma vez que o primeiro objetivo deste trabalho é entendermos características do discurso contendo palavras tóxicas na plataforma Twitter/X, primeiramente, conduzimos uma análise das palavras-chave utilizadas por cada grupo de usuários analisado (i.e., "mais tóxicos" e "menos tóxicos"), e exploramos nuvens de palavras associadas a termos frequentes e representativos de ambos os espectros políticos. Os resultados das referidas análises são apresentados nas seções subsequentes.

Co-ocorrência de Termos. Um fator que deve ser considerado ao realizar análises sobre termos em nossa base de dados, é a co-ocorrência de palavras, visto que, alguns termos existentes no contexto político podem possuir significados diversos no uso popular. Na Tabela 3 são apresentadas as top-10 palavras que mais co-ocorrem com as 10 palavras-chave utilizadas para coleta da base de dados. É evidente a presença de palavras associadas ao contexto político em quase todos os termos buscados, incluindo figuras políticas

[2]A coleta foi efetuada a partir de *tweets* públicos contendo palavras-chave pré-definidas. Logo, existem *tweets* públicos postados por contas configuradas como públicas no momento da coleta. No entanto, durante a coleta das *timelines* essas contas estavam configuradas como privadas, e por este motivo foram desconsideradas deste estudo.

Tabela 3. Top-10 palavras que co-ocorrem algumas palavras-chave.

Palavra-chave	Palavras mais Frequentes (Decrescente)
mortadela	pão (2,1%), comer (1,8%), queijo (1,7%), lula (1,4%), @lulaoficial (1,2%), café (1,2%), reais (1,0%), agora (0,8%), gente (0,7%), dia (0,6%)
petralha	bolsonaro (0,5%), lula (0,4%), cara (0,4%), pt (0,3%), chora (0,3%), brasil (0,3%), @jairbolsonaro (0,2%), termo (0,2%), gente (0,1%), agora (0,1%)
esquerdista	esquerda (8,0%), bolsonaro (6,5%), @nilsonhandebol (6,5%), menos (6,2%), lula (6,0%), nenhum (5,5%), todo (4,0%), todos (2,8%), meta (2,7%), brasil (2,5%)
esquerdopata	bolsonaro (0,6%), brasil (0,5%), @jairbolsonaro (0,5%), presidente (0,5%), contra (0,5%), esquerda (0,4%), sabe (0,3%), ladrão (0,2%), bem (0,2%), lula (0,2%)
comunista	partido (13,6%), brasil (12,5%), lula (11,9%), país (10,5%), comunismo (10,4%), bolsonaro (9,5%), esquerda (7,1%), nunca (6,0%), governo (5,0%), sobre (3,7%)
coxinha	comer (3,9%), pastel (2,8%), frango (2,0%), dia (2,0%), queijo (1,6%), pizza (1,4%), queria (1,3%), hoje (1,3%), coca (1,2%), reais (1,1%)
bolsominion	lula (11,0%), gente (10,6%), cara (8,8%), bolsonaro (8,7%), brasil (8,7%), tudo (7,4%), ainda (5,9%), agora (4,8%), todo (4,7%), ver (4,3%)
fascista	lula (10,5%), bolsonaro (10,5%), gente (9,2%), brasil (8,3%), governo (7,7%), contra (7,7%), ciro (5,7%), fascismo (4,5%), esquerda (4,4%), agora (2,8%)
gado	@gadodecider (20,0%), golpe (17,6%), lula (14,5%), bolsonaro (13,9%), cara (12,9%), agora (10,5%), gente (9,3%), tudo (8,5%), ainda (4,8%), falar (4,6%)
genocida	bolsonaro (12,7%), lula (11,3%), presidente (9,6%), brasil (9,6%), governo (7,5%), pessoas (6,7%), miliciano (5,8%), contra (5,7%), corrupto (4,4%), gente (3,0%)

(como "bolsonaro" e "lula") e também outros adjetivos, como "ladrão", "miliciano" e "corrupto", que são comumente utilizados neste cenário. Porém, existem palavras-chave que também são utilizadas fora do contexto de interesse, como a palavra "coxinha", que apresenta somente termos de culinária co-ocorrentes com ela (como "pastel" e "frango"). A palavra-chave "gado" é a palavra que mais aparece em postagens (ver Figura 2). Observe que o termo que mais co-ocorre com esta palavra chave (i.e., '@gadodecider') que conforme informações contidas na descrição do perfil[3], consiste em um *bot* faz uso de técnicas de inteligência artificial para detectar seguidores do político Bolsonaro, referenciados como gado. Por fim, é importante mencionar que embora este seja um termo bastante presente no contexto político brasileiro, ele pertence a outro tipo de contexto. Dessa forma, um dos desafios da abordagem proposta neste estudo é justamente conseguir distinguir esse tipo de publicação conforme o contexto das palavras utilizadas.

Nuvem de Palavras. A fim de gerar uma análise mais visual do conteúdo frequente da base de dados explorada neste estudo, exploramos nuvens de palavras, que podem resumir grandes volumes de texto, facilitando a identificação dos principais conceitos. As nuvens de palavras das Figuras 3 e 4 foram geradas a partir dos *tweets* postados pelos usuários "mais tóxicos" e pelos "menos tóxicos". É possível notar que, no primeiro grupo, há uma presença mais marcante de palavras relacionadas ao contexto político (como "lula", "bolsonaro" e "presidente"), enquanto no segundo abordagem predominam termos utilizados comumente no cotidiano. Esta análise visual sugere uma correlação entre esses usuários mais tóxicos e o debate político na plataforma Twitter/X.

5. Abordagem para Identificação de Usuários Tóxicos

Nesta seção apresentamos detalhes relativos a abordagem proposta para identificação de usuários tóxicos no contexto do Twitter/X.

[3]`https://x.com/gadodecider`

Figura 3. Nuvem de palavras dos usuários mais tóxicos.

Figura 4. Nuvem de palavras dos usuários menos tóxicos.

5.1. Configurações Experimetais

Pré-processamento dos Dados. Primeiro, foi realizada uma etapa de pré-processamento dos dados, onde eles foram processados para a remoção de links e menções a outros usuários da plataforma. Além disso, também foram removidos *emojis* e algumas marcações próprias do Twitter/X, como "RT". Por fim, foram desconsiderados da análise *tweets* com menos de 4 palavras. Após execução desta primeira etapa, houve uma diminuição de 103.595 e 134.099 *tweets* das bases de dados de usuários "mais tóxicos" e "menos tóxicos", respectivamente. É importante destacar também que a biblioteca NLTK[4] foi utilizada para remoção de *stop-words*. Ao final dessas duas etapas, a base de dados de *tweets* postados por usuários "mais tóxicos" é composta por 454.643 publicações, enquanto que a base de dados de usuários "menos tóxicos" 465.035 *tweets*. A próxima etapa, de preparação dos dados, envolve a organização dos *tweets* em vetores dentro de um dicionário, utilizando como índice o ID do autor daquela publicação. Em uma análise manual, identificamos que para alguns autores, principalmente os que possuíam um grande volume de postagens, havia a existência de textos repetidos, com caráter de *spam/bots*, por esse motivo, foi realizada a remoção de *tweets* idênticos (similaridade = 1).

Representação Vetorial dos Usuários. As entradas utilizadas foram obtidas a partir do BERTimbau[5], um modelo de processamento de linguagem natural BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2018] pré-treinado com frases em português que possui a capacidade de capturar o contexto das palavras em uma frase. Os vetores obtidos a partir do modelo possuem 29.794 posições para cada palavra na entrada. O BERT é um modelo de processamento de linguagem natural desenvolvido pelo Google. Ele utiliza a arquitetura Transformer para capturar relações contextuais entre palavras em uma sentença. Diferente de modelos anteriores, o BERT adota uma abordagem bidirecional, considerando palavras anteriores e posteriores para compreender o contexto. Isso é possível devido ao uso de mecanismos de atenção na arquitetura Transformer.

Contudo, como o BERTimbau retorna um vetor para cada palavra, propusemos um esquema para representação dos dados (i.e., *timeline* de um usuário fornecido como entrada) composto por três etapas principais. Uma visão geral é apresentada na Figura 5: (1) Para todas as postagens realizadas por cada usuário classificado como "mais tóxico" ou "menos tóxico", foram obtidas as representações vetoriais utilizando o BERTimbau para cada palavra em cada uma das publicações; (2) Para gerar as representações das

[4] https://www.nltk.org/

[5] https://huggingface.co/neuralmind/bert-base-portuguese-cased

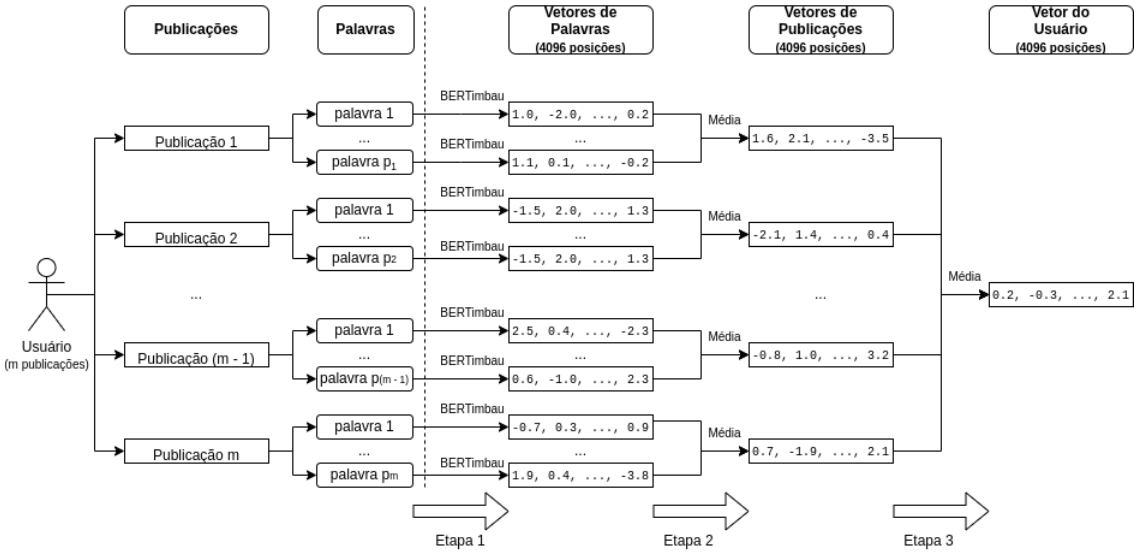

Figura 5. Esquema de representação dos usuários a partir das suas publicações.

publicações foi realizada a média, elemento a elemento, dos vetores das palavras presentes nos *tweets*, e, por fim; (3) Para gerar as representações dos usuários foi realizada a mesma operação anterior, agora entre as representações das publicações de cada usuário. Se tratando da classificação de usuários com quantidades diferentes de *tweets*, foi importante determinar a descrição do vetor de características com dimensão invariante ao número de postagens.

Classificadores e Detalhes. Para a seleção do modelo de classificação foi realizado um experimento inicial com validação cruzada com 5 partições utilizando o conjunto de dados do "Primeiro turno" das eleições que compreende o período de 02/01 a 02/10 de 2024. Os testes foram executados com 5 abordagens de aprendizado de máquina distintas, sendo elas: *Multi-layer Perceptron Classifier* (MLP), *Linear Support Vector Classification* (Linear SVC), *C-Support Vector Classification* (SVC), *Random Forest Classifier* (RF) e, *Gradient Boosting Classifier* (XGB). Todos os modelos foram testados utilizando os parâmetros padrões disponibilizados pela biblioteca *scikit-learn 1.2.2*[6]. É importante destacar que durante este experimento não foi realizada otimização de parâmetros de nenhum classificador, pois sem um estudo específico poderia ser injusto comparar a seleção para cada um dos modelos. Os resultados são apresentados e discutidos a seguir.

5.2. Resultados de Classificação

A Tabela 4 apresenta a média e o desvio padrão das métricas *F1 score* e *Compute Area Under the Receiver Operating Characteristic Curve* (ROC AUC) nas 5 partições. Estas são métricas amplamente utilizadas em tarefas de classificação [Baeza-Yates et al. 1999]. *F1 score* pode ser considerada como uma métrica de equilíbrio entre precisão e revocação, considerando tanto falsos positivos como falsos negativos em um único valor. Já a métrica ROC AUC calcula a área sob a curva *ROC*, avaliando como o modelo se comporta com diferentes valores de para o limiar de decisão, capturando a compensação entre taxa de verdadeiros positivos (resultado esperado) e taxa de falsos positivos (resultado não desejado). Para ambas um valor mais alto representa um desempenho melhor da abordagem

[6] `https://scikit-learn.org/1.2/`

Tabela 4. Desempenho dos classificadores com validação cruzada

Classificador	*F1 score*	ROC AUC
MLP	0,41±0,00	0,50±0,00
SVC	0,79±0,04	0,89±0,03
Linear SVC	0,80±0,07	0,93±0,01
RF	0,83±0,03	0,93±0,02
XGB	**0,85±0,01**	**0,94±0,01**

a ser avaliada. As métricas foram escolhidas por se tratar de uma base de dados desbalanceada, nesse sentido, considerando a expectativa (talvez otimista) de que a maioria dos usuários não é tóxica, temos uma base mais similar à realidade. Podemos observar que o classificador XGB obteve os melhores valores para ambas as métricas. Especificamente, 0,85 e 0,94 para *F1 score* e ROC AUC, respectivamente. Por outro lado, notamos um resultado significativamente inferior em ambas as métricas para o MLP, com cerca de 48% do melhor valor encontrado na métrica *F1 score*. É importante destacar que todos os outros modelos possuem interseção nas métricas considerando os desvios-padrão, o que indica um empate estatístico. No entanto, em função dos resultados obtidos, o XGB foi selecionado como melhor modelo durante esta etapa.

Depois, realizamos um segundo experimento para verificar o comportamento do modelo utilizando diferentes repartições do dado, ou seja, temporalidade mensal. Os resultados são apresentados na Tabela 5. O objetivo desse teste foi avaliar o quanto as métricas obtidas podem ser beneficiadas e/ou prejudicadas utilizando um período de coleta diferente, com dados considerando um período de tempo mais curto, podendo fornecer um indicativo de aplicabilidade da abordagem em um cenário real. Notamos que, mesmo considerando apenas os dados de 1 mês de coleta, a média obtida para a métrica *F1 score*, por exemplo, se manteve em cerca de 97% da média ótima do experimento anterior, sendo assim pode-se afirmar que o experimento realizado apresenta evidências de que o modelo (i.e., RF) mantém bons resultados mesmo com períodos curtos de coleta de dados, o que pode viabilizar a aplicação em um cenário real. Dessa maneira, é plausível considerar a utilização do modelo proposto mesmo com um conjunto de dados relativamente pequeno sem obter um prejuízo significativo nos resultados, tornando-o mais considerável para aplicações práticas.

5.3. Análise do Potencial Prático da Abordagem Proposta

Na seção anterior, descrevemos um modelo proposto para a identificação de usuários tóxicos no Twitter/X, baseado nas postagens efetuadas por eles durante o "Primeiro turno"

Tabela 5. Avaliação do XGB com diferentes subconjuntos dos dados

Intervalo do subconjunto	*F1 score*	ROC AUC
02/04 a 02/10	0,85 ± 0,02	0,93 ± 0,01
02/05 a 02/10	0,85 ± 0,03	0,93 ± 0,01
02/06 a 02/10	0,85 ± 0,03	0,93 ± 0,01
02/07 a 02/10	0,84 ± 0,03	0,92 ± 0,02
02/08 a 02/10	0,84 ± 0,01	0,93 ± 0,02
02/09 a 02/10	0,83 ± 0,02	0,92 ± 0,02

Tabela 6. Amostra de *tweets* na base de dados.

Tweet
"lula é ladrâo rle vai dizer wye foi para tomar uma cervejinha #bolsonaroreeleito"
"carla zambelli é uma vigarista. sempre foi. mentirosa compulsiva. propagadora de mentiras. assediadora. a casa dela vai cair assim como a de sua parceira de crimes, damares. é estarrecedor que ambas sigam impunes depois de tudo que fizeram."
"você vota em marginal e somos nós que badernamos? vão se ferrar você com seu ladrão, sua."
"o anti cristo pai da mentira e seus seguidores agora estão aterrorizando até a polícia acorda povo de deus."
"compoamento do bozo com meninas venezuelanas é sim pedófilo. e ainda tem evangélicos que vota nesse canalha."
"de todas as maldades do bozo a pior é ter levado seus 04 filhos e mulheres a roubar. . . "
"tu é um bostinha de classe média esquerdista liberal, típico eleitor do psol."
"é mentira, como tudo dessa extrema direita assassina."

das eleições presidenciais brasileiras de 2022. A partir disso, utilizamos a base de dados do "Segundo turno" das eleições para investigar o potencial impacto prático da abordagem proposta. Especificamente, conforme mencionado na Seção 3, nossa base de dados é composta por um total 367 usuários potencialmente tóxicos. Todos esses usuários, em conjunto, são responsáveis por um total de 138.821 *tweets* postados no período do "Segundo turno". Desses 367 classificados como tóxicos, o modelo proposto na seção anterior foi capaz de identificar 85% (ou seja, 313 usuários tóxicos), que foram responsáveis pela postagem de 103.677 *tweets* (do total de 138.821). Em outras palavras, isso significa que, o modelo, se aplicado em tempo hábil neste contexto poderia ter evitado um total de cerca 100 mil postagens potencialmente tóxicas na plataforma durante o "Segundo turno" das eleições. Isso inclui, por exemplo, a aplicação de medidas de restrição, que podem incluir mensagens de aviso, banimento da plataforma ou, em ultimo grau, sendo passível o encaminhamento às autoridades legais, aumentando a responsabilização dos usuários e diminuindo a sensação de impunidade causada pelo aparente anonimidade. Vale ressaltar que o impacto desse tipo de conteúdo, especialmente em plataformas sociais, pode incitar a polarização política e incentivar a replicação desse tipo de atitude, dado o contexto e as especifidades causadas por esse tipo de ambiente.

Por fim, para ilustrar os tipos de *tweets* que poderiam ter sido evitados, selecionamos uma amostra dos dados, como pode ser visto na Tabela 6, proporcionando uma perspectiva qualitativa dessas mensagens do contexto político. A análise realizada visa fornecer uma visão geral de como a grande parte desses usuários, estão de fato inseridos no contexto da discussão política das eleições nas plataformas sociais. Entre as palavras mais frequentes no conjunto de dados analisado, como podemos ver na Figura 6, notamos que "Bolsonaro" aparece com destaque, com 13.281 citações, seguido por "Lula" em quarto lugar, com 10.537 aparições. Essas palavras ganham destaque na nuvem de palavras gerada com todos os *tweets* analisados. Esses dados indicam a relevância dos temas ligados aos candidatos, bem como expressões que refletem opiniões e engajamento político por parte dos usuários no Twitter/X durante o período analisado.

6. Conclusão e Trabalhos Futuros

Neste trabalho propusemos uma abordagem para identificar usuários (potencialmente) tóxicos na plataforma do Twitter/X durante as eleições presidenciais de 2022. Para isso, a partir de um conjunto pré-definido de palavras-tóxicas coletamos 6.9 milhões de *tweets*

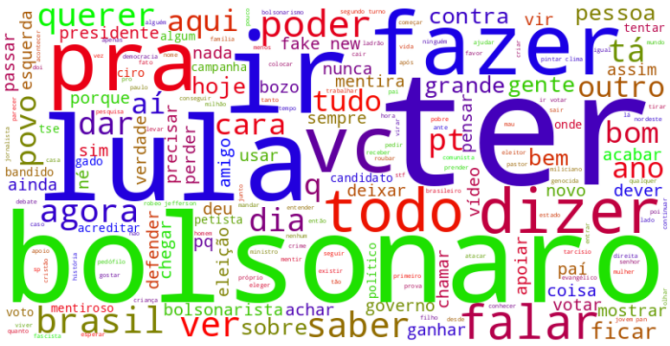

Figura 6. Nuvem de palavras com os termos mais utilizados entre os usuários tóxicos durante o período do segundo turno.

de aproximadamente 1.7 milhão de usuários. Esses usuários foram ordenados com base na frequência de utilização de palavras-chave que podem desencadear discussões tóxicas e o resultado desta etapa foi utilizado para separação deles grupos de "mais tóxicos" ou "menos tóxicos". Nossos resultados revelam várias descobertas interessantes.

Nossa caracterização dos dados, revelaram padrões interessantes de comunicação neste contexto que podem ser utilizados para distinguir usuários potencialmente tóxicos dos demais. A utilização do modelo de processamento de linguagem natural pré-treinado em português, BERTimbau, nos permitiu gerar representações que capturaram o contexto dos textos, possibilitando a utilização dos dados para treinar nossos modelos. Em destaque, as abordagens investigadas obtiveram resultados promissores na classificação de usuários tóxicos, alcançando métricas satisfatórias em sua avaliação, consolidando-se como escolhas viáveis como ferramentas de suporte as plataformas de mídias sociais neste contexto.

Além disso, foi possível identificar os usuários classificados como potencialmente tóxicos pelo nosso modelo, e realizar uma análise comportamental dos mesmos no período do segundo turno. Os resultados evidenciaram que, se a plataforma tivesse agido e interceptado as ações desses usuários anteriormente a esse período, seria possível conter danos significativos durante a fase decisiva do processo eleitoral A detecção prévia de comportamentos tóxicos poderia ter contribuído para um ambiente online mais saudável e promovido um debate político mais construtivo. Portanto, a implementação de estratégias de moderação baseadas em modelos preditivos demonstra ser uma abordagem promissora para contribuir para garantia e integridade do espaço digital durante eventos políticos e sociais cruciais, como os turnos de uma eleição presidencial.

Por fim, acreditamos que a abordagem proposta neste estudo para identificação de usuários tóxicos possa ser adaptada para outros contextos e/ou plataformas. Logo, como trabalhos futuros, pretendemos explorar outras bases de dados. Além disso, planejamos explorar diferentes abordagens de modelagem e outras técnicas de processamento de linguagem natural para refinar ainda mais a precisão das classificações de usuário tóxico.

Agradecimentos

Este trabalho foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Universidade Federal de Viçosa (UFV).

Referências

Al-Hassan, A. and Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In *International Conference on Computer Science and Information Technology*, volume 10, pages 10–5121.

Almerekhi, H., Kwak, H., Salminen, J., and Jansen, B. J. (2020). Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference (WWW)*, page 3033–3040.

An, J., Kwak, H., Lee, C. S., Jun, B., and Ahn, Y.-Y. (2021). Predicting anti-asian hateful users on twitter during covid-19. In *Findings of the Association for Computational Linguistics (EMNLP)*, page 4655–4666.

Araujo, M. M., Ferreira, C. H., Reis, J. C., Silva, A. P., and Almeida, J. M. (2023). Identificação e caracterização de campanhas de propagandas eleitorais antecipadas brasileiras no twitter. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BrasNAM)*, pages 67–78.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Christhie, W., Reis, J. C. S., Moro, M. M., Benevenuto, F., and Almeida, V. (2018). Detecção de posicionamento em tweets sobre política no contexto brasileiro. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BrasNAM)*.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Proc. of the Int'l Conference on Web and Social Media*, pages 89–96.

da Fonseca, L. G. G., Ferreira, C. H., and Reis, J. C. S. (2024). The role of news source certification in shaping tweet content: Textual and dissemination patterns in brazil's 2022 elections. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guimaraes, S., Silva, M., Caetano, J., Araújo, M., Santos, J., Reis, J. C. S., Silva, A. P., Benevenuto, F., and Almeida, J. M. (2022). Análise de propagandas eleitorais antecipadas no twitter. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Kertzman, R. (2020). Na guerra de fake news, quem mente mais: bolsominions ou petralhas? https://www.em.com.br/app/colunistas/ricardo-kertzman/2022/02/15/interna_ricardo_kertzman,1345242/amp.html.

Lima, L., Reis, J. C., Melo, P., Murai, F., and Benevenuto, F. (2020). Characterizing (un) moderated textual data in social systems. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 430–434.

Neves, M. (2020). Direita x esquerda: 12 nomes do futebol que nunca ficaram em cima do muro!... https://www.uol.com.br/esporte/colunas/milton-neves/2022/08/27/direita-x-esquerda-12-nomes-do-futebol-que-nunca-ficaram-em-cima-do-muro.htm.

Queiroga, L. (2022). Casos de homicídio por motivação política marcaram reta final da eleição. https://extra.globo.com/noticias/brasil/eleicoes-2022/casos-de-homicidio-por-motivacao-politica-marcaram-reta-final-da-eleicao-relembre-25582071.html.

Reis, J. C., Melo, P., Belém, F., Murai, F., Almeida, J. M., and Benevenuto, F. (2023a). Helping fact-checkers identify fake news stories shared through images on whatsapp. In *Proc. of the Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 159–167.

Reis, J. C., Melo, P., Silva, M., and Benevenuto, F. (2023b). Desinformação em plataformas digitais: Conceitos, abordagens tecnológicas e desafios. *Jornada de Atualiação em Informática (JAI). Sociedade Brasileira de Computação (SBC)*.

Silva, F. and Freitas, L. (2022). Brazilian portuguese hate speech classification using bertimbau. In *The International FLAIRS Conference Proceedings*, volume 35.

Teixeira, M. C. and Reis, J. C. (2023). Análise do discurso de ódio em comentários de vídeos no youtube: Um estudo de caso da cpi da covid-19 no brasil. In *Anais doSimpósio Brasileiro de Bancos de Dados (SBBD)*, pages 330–335.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., and Blackburn, J. (2018). What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014.