V. 16 – 2025
ISSN: 2179-1465 / https://www.revistageminis.ufscar.br
DOI: 10.14244/2179-1465.RG.2025v16p275-296

DISCURSOS DE ÓDIO EM REDES SOCIAIS: UMA ANÁLISE COM PROCESSAMENTO DE LINGUAGEM NATURAL

HATE SPEECH ON SOCIAL MEDIA: AN ANALYSIS WITH NATURAL LANGUAGE PROCESSING

DISCURSOS DE ODIO EN LAS REDES SOCIALES: UN ANÁLISIS CON PROCESAMIENTO DEL LENGUAJE NATURAL

Gustavo Gurjão Camargo Campos
Universidade Federal de Campina Grande, CEEI, UASC
ORCID: https://orcid.org/0009-0009-2651-6682
Campina Grande, Paraíba, Brasil

Eanes Torres Pereira
Universidade Federal de Campina Grande, CEEI, UASC
ORCID: https://orcid.org/0000-0002-9717-794X
Campina Grande, Paraíba, Brasil

Sylvia Iasulaitis
Universidade Federal de São Carlos
ORCID: https://orcid.org/0000-0002-3526-1003
São Carlos, São Paulo, Brasil

Recebido: 15/03/2025 / Aprovado: 03/08/2025
Como citar: CAMPOS, G. G. C.; PEREIRA, E. T., IASULAITIS, S. Discursos de Ódio em Redes Sociais: uma análise com processamento de linguagem natural. Revista GEMInIS, v. 16, p. 275-296, 2025

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

RESUMO

Houve um aumento expressivo na disseminação de discursos de ódio e preconceituosos em sites de redes sociais durante os pleitos eleitorais, sendo o maior deles de teor xenófobo, cujo crescimento em 2022 foi de 821%. O objetivo deste trabalho foi investigar, por meio do Processamento de Linguagem Natural, a variação das opiniões acerca do Nordeste no Twitter (X) no ano de 2022. Foi possível identificar que a opinião em relação aos nordestinos foi eminentemente negativa e os discursos de ódio aumentaram à medida em que as eleições se aproximaram, cujos conteúdos se encaixam dentro do espectro da xenofobia.

Palavras-chave: Redes sociais; Discurso de ódio; Processamento de linguagem natural.

ABSTRACT

There was a significant increase in the dissemination of hate speech and prejudice on social media sites during the elections, the largest of which was xenophobic, with an increase of 821% in 2022. The objective of this work is to investigate, through Natural Language Processing, the variation in opinions about the Northeast on Twitter (X) in 2022. It was possible to identify that the opinion towards people from the Northeast was eminently negative and hate speech increased as the elections approached, the contents of which fit within the spectrum of xenophobia.

Keywords: Social media; Hate speech; Natural language processing.

RESUMEN

Hubo un aumento significativo en la difusión de discursos de odio y prejuicios en las redes sociales durante las elecciones electorales, el mayor de ellos de carácter xenófobo, cuyo crecimiento en 2022 fue del 821%. El objetivo de este trabajo es investigar, a través del Procesamiento del Lenguaje Natural, la variación de opiniones sobre el Nordeste en Twitter (X) en el año 2022. Fue posible identificar que la opinión en relación a los nordestinos era eminentemente negativa y los discursos de odio aumentaron a medida que se acercaban las elecciones, cuyos contenidos encajan dentro del espectro de la xenofobia.

Palabras Clave: Redes sociales; Discurso de odio; Procesamiento del lenguaje natural.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

1. INTRODUÇÃO

Os sites de redes sociais se tornaram centrais para as interações durante as campanhas eleitorais, devido à sua acessibilidade e facilidade de uso (Barberá & Rivero, 2015). Contudo, é possível identificar que houve um aumento expressivo na disseminação de comentários ofensivos e discursos de ódio nas plataformas de mídias sociais, em especial durante os pleitos eleitorais.

De acordo com a ONG Safernet, que opera em cooperação com o Ministério Público Federal, 2022 foi o terceiro ano eleitoral consecutivo com crescimento de crimes de ódio, dentre eles: xenofobia, intolerância religiosa, misoginia, apologia a crimes contra a vida, LGBTQIA+fobia, neonazismo e racismo.

Especificamente, conforme pode ser verificado na Tabela 1, o maior aumento registrado durante o pleito de 2022 foi em relação à xenofobia, que cresceu 821%; em segundo lugar estiveram os crimes de intolerância religiosa, com aumento de 522% e em terceiro lugar os de misoginia, com crescimento de 184%[1].

Tabela 1 – Denúncias de crimes de ódio aumentam em anos eleitorais

Crimes de ódio	2017	2018*	Aumento	2019	2020*	Aumento	2021	2022	Aumento
Apologia a crimes contra a vida	10611	27713	161,17%	8182	11852	44,85%	7390	7623	18,20%
LGBTFobia	2592	4244	63,73%	2752	5293	92,30%	5347	6829	51,90%
Misoginia	961	16717	1639,50%	7112	12698	78,50%	8174	20924	184,00%
Neonazismo	1172	4244	262,10%	1071	9004	740,70%	14476	2033	-85,40%
Racismo	6166	8336	35,10%	4310	10684	147,80%	6888	6586	14,30%
Xenofobia	1395	9703	595,50%	978	2066	111,20%	1097	7075	**821%**
Intolerância religiosa	1459	1084	-25,70%	1413	1321	-6,51%	759	3818	522%
Total de denúncias	24356	72041	195,78%	25818	52918	104,96%	44131	54888	39,30%

Fonte: Dados da ONG Safernet e Ministério Público Federal

[1] Tabela de dados disponível em https://docs.google.com/spreadsheets/d/1UXTafDanz6hszfAFbOmnO5dxsd6_cMX1Mi7OimdjtWU/edit#gid=0

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Em 2022, o descontentamento com o resultado das eleições, devido à ampla votação de Lula no Nordeste, levou novamente a uma explosão de denúncias de xenofobia, fato que já havia sido registrado em 2014, quando Dilma Rousseff venceu Aécio Neves e em 2018 quando a região registrou boa votação em Haddad com relação à Bolsonaro.

Os resultados das eleições presidenciais de 2022 determinaram a vitória do candidato do PT Luís Inácio "Lula" da Silva - político nordestino oriundo de Garanhuns, município pernambucano - sobre o então presidente e candidato à reeleição Jair Bolsonaro. Analisando a geografia do voto, a nível estadual Lula garantiu vitória em 13 dos 26 estados brasileiros, sendo 9 deles: Rio Grande do Norte, Paraíba, Pernambuco, Sergipe, Alagoas, Ceará, Bahia, Maranhão e Piauí, no caso todos os estados do Nordeste[2].

Devido a esta geografia do voto, o Nordeste surgiu para o senso comum como sendo o responsável pelo resultado das eleições, o que gerou uma grande variedade de comentários acerca da região e de seus habitantes nos sites de redes sociais.

O Twitter, atualmente denominado X, é uma das plataformas de redes sociais mais utilizadas no Brasil[3]. Conhecido por sua política de postagens curtas e rápidas, é um meio pelo qual as pessoas podem facilmente e rapidamente postar suas opiniões imediatas acerca de qualquer tópico. Por conta do curto tamanho dos textos e uma tendência à postagem de opiniões e pensamentos, o Twitter é um site de rede social propício para análises textuais.

O objetivo deste trabalho é investigar a variação das opiniões dos usuários do Twitter (X) acerca do Nordeste, visando identificar se as eleições presidenciais de 2022 exerceram influência neste sentido. O intuito é analisar a evolução do destaque dado à região na rede social à medida em que os meses de votação se aproximaram e a variação da opinião de seus usuários sobre os nordestinos durante esse período.

Com Processamento de Linguagem Natural, buscou-se associar o termo "nordestino" e suas variações a diferentes palavras e avaliar como essa associação mudou com o passar dos meses, especialmente focando em saber se palavras pejorativas tiveram sua associação aumentada, diminuída ou mantida estável quanto mais se aproximava o pleito eleitoral.

[2] https://infograficos.oglobo.globo.com/politica/eleicoes-2022/mapa-votacao-municipios-e-estados-do-brasil.html#/presidente?desempenho=geral
[3] https://www.techtudo.com.br/noticias/2019/02/conheca-as-redes-sociais-mais-usadas-no-brasil-e-no-mundo-em-2018.ghtml

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

2. DISCURSOS XENÓFOBOS

O foco deste trabalho será o discurso xenófobo dirigido aos nordestinos. De acordo com Serrão (2020), a linguagem xenófoba dirigida por usuários de redes sociais contra os nordestinos faz parte de um *continuum* histórico de opressão fomentado por estereótipos regionais.

Os discursos de ódio e preconceito se assentam em visões que definem o Nordeste e seu povo como sendo "atrasados", o que reflete a formação racial e o preconceito regional que caracterizam o Brasil desde o final do século XIX. O preconceito regional no Brasil, especificamente contra o nordestino, se constitui em uma forma de xenofobia moderna (Serrão, 2022) e mostra a centralidade de premissas estereotipadas sobre cultura, raça e classe socioeconômica. De acordo com Albuquerque Jr. (1999, p. 307) "o Nordeste, assim como o Brasil, não são recortes naturais, políticos ou econômicos apenas, mas, principalmente, construções imagético-discursivas, constelações de sentido".

Muitas vezes, os residentes dos nove estados brasileiros que compõem o Nordeste são considerados "subnacionais, súditos coloniais, um povo inferior, um grupo de pessoas que não pertence ao Brasil moderno" (Grosfoguel, 2003; Quijano, 2005 apud Serrão, 2022).

As representações sociais xenófobas atacam a dignidade de todo um grupo social que compartilha a característica ensejadora da discriminação, o que se denomina vitimização difusa. Embora não se possa distinguir quem, nominal e numericamente, são as vítimas, existem inúmeras pessoas atingidas devido a seu pertencimento a esse grupo social (Martins, 2019).

Ao instigar e incitar preconceito ou ódio, o emissor pretende angariar adesão, principalmente fazendo uso de estratégias comunicacionais. Por contar com poucos filtros ou com moderação limitada, os sites de redes sociais têm sido amplamente utilizados para tanto, por serem compreendidos como espaços para emissão de "opinião", de "liberdade de expressão" e legitimados por um peculiar senso de propriedade: "minha conta", "meu mural", "minha página". Assim, na ambiência das plataformas de mídias sociais, o discurso preconceituoso é comumente legitimado como opinião proprietária que deve ser permitida e que não se caracteriza como ofensa pública (Hoepfner, 2014).

Desde 2010 circulam centenas de postagens insultando e ameaçando os nordestinos, vistos como responsáveis por levar e manter o PT no poder. Em 2010, após as eleições que consagraram Dilma Rousseff do Partido dos Trabalhadores (PT) presidenta da República, diversas manifestações de ódio de cunho xenófobo foram publicadas no Twitter, como ilustra a Figura 1.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Figura 1 – Exemplos de discurso de ódio contra nordestinos em campanhas eleitorais no Brasil

Fonte: Twitter/X

Serrão (2022), estudando as publicações discriminatórias nas mídias sociais após as eleições presidenciais de 2014 e 2018 no Brasil, revela semelhanças na maior parte da linguagem racista e xenofóbica nos dois ciclos eleitorais, mas um aumento a animosidade em relação ao Partido dos Trabalhadores em 2018, dado que foi atribuído aos nordestinos o papel de impedir a vitória de Bolsonaro no primeiro turno.

O ex-presidente de extrema-direita, Jair Bolsonaro, durante suas campanhas eleitorais, acentuou a xenofobia e o sentimento de ódio contra os nordestinos, ao disseminar estereótipos e preconceito em seus vídeos nas redes sociais. Em entrevista concedida à Record News em 2012, afirmou que "*Bolsa família é uma mentira, no Nordeste você não consegue uma pessoa pra trabalhar na tua casa*"; em outubro de 2014 declarou: "*Você vê meninas no Nordeste, batem a mão na barriga grávida e fala o seguinte – tem também o auxílio natalidade – 'esse aqui vai ser uma geladeira', 'esse aqui vai ser uma máquina de lavar', e não querem trabalhar!*", e, ainda, em uma das lives realizadas em 2016, afirmou: "*Tem um cearense... um cabeçudo aqui do meu lado, pô, eu acho que o estômago é maior que a cabeça dele*" (Iasulaitis & Vicari, 2021).

Além de tratar aspectos físicos de forma pejorativa e relacionar a região a políticas denominadas "assistencialistas", as críticas online associam também a seca e a pobreza no Nordeste como "castigo", bem como a migração interna para o Sul e Sudeste.

3. MÉTODOS E PROCEDIMENTOS DE PESQUISA

3.1 Processamento de Linguagem Natural

As técnicas de Aprendizagem de Máquina e o Processamento de Linguagem Natural têm sido utilizadas para tarefas de classificação e detecção de discurso de ódio (Mullah & Zainon, 2021). Oluwafemi & Kotze (2020) realizaram estudos de detecção de linguagem de ódio no Twitter no contexto da África do Sul, utilizando técnicas de Aprendizagem de Máquina e de classificação para

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

detectar e classificar *tweets* que continham esse tipo de conteúdo. No Brasil, Araújo *et al* (2020) realizaram estudos de dados do Twitter, envolvendo análise exploratória de dados, analisando textos e *hashtags*. Barbosa (2021) utilizou dados coletados do Twitter para construir uma rede neural com o intuito de detectar linguagem transfóbica.

O Processamento de Linguagem Natural - (PLN) é uma área da Inteligência Artificial que tem o objetivo de utilizar os computadores para trabalhar com um grande volume de dados não passível de ser analisado manualmente, com a finalidade de que possam entender e interpretar a linguagem humana. Um dos conceitos da área de PLN é o de pré-processamento textual, que se refere ao processo de transformação de um texto de seu estado original de coleta para outro, em que possa ser analisado, visando à realização da tarefa-alvo (Liddy, 2001).

3.2 Procedimentos de Pesquisa

Para a realização desta pesquisa, foram seguidos os seguintes passos: coleta de dados, pré-processamento desses dados e a realização de quatro tipos de análises, com diferentes métodos, com a intenção de identificar a variação da opinião dos usuários do *Twitter* acerca do povo nordestino no decorrer do período pré-eleitoral e eleitoral no ano de 2022.

A principal decisão tomada nesta pesquisa quanto às palavras e termos analisados foi que as palavras sementes para a busca seriam variações da palavra nordestino. Especificamente, as formas-raiz “nord” (forma-raiz de “nordeste”) e “nordestin” (forma-raiz de nordestino, nordestina, nordestinos, nordestinas e afins). Como será melhor detalhado nas subseções seguintes, a partir dessas formas-raiz foram realizadas análises para identificar palavras e termos que surgiram nos dados coletados e que estavam correlacionadas com elas. Esta decisão é justificada pelo fato de que seria impraticável listar todos os termos e palavras que podem ser associados a nordestinos. Além disso, termos que supostamente poderiam estar correlacionados com variações da palavra nordestino poderiam também estar correlacionados a outros contextos. Usar palavras que podem estar associadas a múltiplos contextos abriria demasiadamente o escopo de análise e poderia causar trabalho desnecessário.

3.2.1 Coleta de dados

Para a coleta dos dados, foram criadas contas no *Developer Twitter*, com a intenção de ter acesso à API do Twitter. Com a conta criada e o acesso garantido, foram feitas requisições à API, buscando *tweets* que continham os nomes e apelidos dos dois principais candidatos à Presidência da

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

República, Lula e Bolsonaro. Esses *tweets* foram coletados por meio da API do Twitter, usando como filtro palavras-chave. O período de análise dos *tweets* coletados foi entre os meses de julho e dezembro.

Uma possível limitação da base de dados utilizada, que não invalida a presente pesquisa, é que ela foi coletada durante o período de pleito eleitoral ou poucos meses após as eleições de 2022. Isso se configura como limitação, pois não foram coletados dados do ano inteiro e no período de coleta havia maior ocorrência de discursos xenófobos nas redes sociais, especialmente na rede analisada, o X (Twitter). Outro fator que deve ser considerado é que a base de dados foi construída utilizando termos de busca do contexto das eleições presidenciais e não necessariamente com o objetivo principal de avaliar xenofobia. No entanto, os dados coletados continham discurso xenofóbico e, portanto, uma amostra deles foi analisada por essa perspectiva. Quanto aos detalhes sobre os termos, os filtros e as palavras-chave utilizadas para construir a base de dados (Iasulaitis et al, 2025), eles podem ser melhor entendidos a partir da documentação e publicações de divulgação criadas pelo grupo de pesquisa Interfaces.

Assim, os dados utilizados nesta pesquisa são uma amostra da coleta feita pelo grupo de pesquisa do qual os autores fazem parte. O *dataset* denominado *The Interfaces Twitter Elections Dataset*, conhecido pelo acrônimo ITED-Br, é composto por 282 milhões de *tweets* e é a terceira maior base de dados de *tweets* com propósitos políticos do mundo (Iasulaitis et al, 2025).

3.2.2 Pré-Processamento

Neste trabalho, foram utilizadas algumas técnicas necessárias para a realização do pré-processamento dos dados (Vijayarani, 2015), a saber:

- *Lowercasing*: técnica que consiste em transformar todas as letras de um texto em letras minúsculas, com o objetivo de garantir que a mesma palavra, escrita com letras minúsculas e maiúsculas em diferentes partes do texto, seja analisada igualmente. Por exemplo, o *lowercasing* da frase ‘‘Nordestinos são incríveis’’ teria como resultado ‘‘nordestinos são incríveis’’.
- Tokenização (*tokenization*): técnica que consiste em dividir o texto em uma lista de palavras ou *tokens* individuais. Por exemplo, a tokenização da frase ‘‘nordestinos são incríveis’’ resultaria nos tokens ‘‘nordestinos’’, ‘‘são’’ e ‘‘incríveis’’.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

- Stemização (*stemming*): técnica que consiste em retirar as inflexões da palavra, reduzindo-a à sua forma raiz. Por exemplo, a stemização do *token* ‘‘nordestinos’’ resultaria no *token* ‘‘nordestin’’.
- Remoção de caracteres especiais: técnica que consiste na remoção dos caracteres especiais de um texto ou *token.*
- Remoção de *stopwords*: remoção de palavras com pouquíssimo significado semântico, como artigos e preposições. Exemplos de *stopwords* são: “a”, “o” e “de”.

3.2.3 Análise Exploratória dos Dados

A análise exploratória de dados textuais se refere ao conjunto de técnicas usadas para analisar e investigar conjuntos de dados, resumir suas principais características, investigar padrões encontrados, testar hipóteses e facilitar sua visualização e leitura (Martinez *et al*., 2017).

As técnicas e algoritmos usados para analisar os dados nesta pesquisa foram:

1. *Bag of Words:* representação de um conjunto de dados de acordo com a ocorrência de cada palavra ou *token* encontrado nele.
2. *Distribuição de Frequência*: técnica que mapeia *tokens* com sua frequência em um texto. Com ela, também é possível prever uma associação entre termos ou *tokens* presentes em um conjunto de dados com base em suas respectivas frequências.
3. *Word Embeddings*: técnica em que palavras são representadas como vetores de valores reais em um espaço dimensional. Nessa situação, vetores semanticamente similares tendem a estar perto uns dos outros. Ou seja, palavras que são graficamente colocadas próximas umas das outras em um plano são semelhantes semanticamente.
4. *Word2Vec*: algoritmo baseado em vetores calculados em *Word Embeddings*, que busca mapear palavras ou *tokens* com significados semânticos associados. Essa associação é calculada a partir da ideia de que palavras que frequentemente são acompanhadas das mesmas palavras vizinhas tendem a ser semanticamente associadas. Esse algoritmo classifica uma associação entre dois termos de 0% (nenhuma associação) a 100% (significado semântico igual).

4. RESULTADOS E DISCUSSÃO

Nesta seção, serão apresentados os resultados de cada etapa da pesquisa, realizadas com as técnicas e algoritmos anteriormente apresentados.

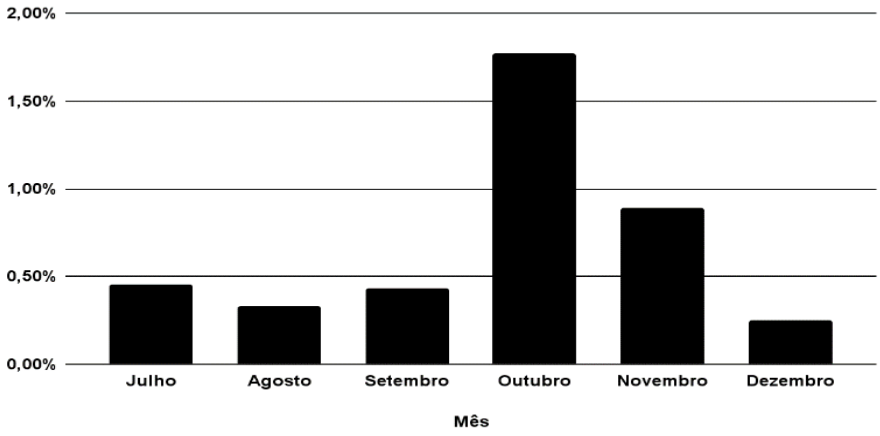
Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

4.1 Presença nos Tweets

A primeira etapa da análise foi verificar a presença de palavras e termos que se referem a nordestinos nos dados coletados e a variação dessa presença no decorrer dos meses de votação. Para isso, foi feita uma análise numérica simples. Para cada mês, foram comparadas as porcentagens de postagens que possuíam as formas-raiz "nord" (forma-raiz de "nordeste") e "nordestin" (forma-raiz de nordestino, nordestina, nordestinos, nordestinas e afins) em relação ao total de *tweets*. Assim, foi possível observar em quais meses os nordestinos tiveram maior presença nos *tweets* e o impacto em comparação aos outros meses.

Como é possível verificar na Figura 2, a proporção de postagens que citaram o Nordeste aproximadamente triplicou no mês de outubro, o que mostra um aumento significativo da presença de conversas acerca do Nordeste no mês logo após o primeiro turno e quando ocorreu o segundo turno das eleições. Sabendo-se com antecedência que em períodos eleitorais pode haver um aumento de discurso de ódio contra pessoas de regiões diferentes do país, podem ser criadas medidas educativas e de coibição de tais conteúdos com antecedência. Essas ações podem ser realizadas tanto pelos órgãos governamentais competentes quanto pelas próprias empresas proprietárias das plataformas de redes sociais. Sabe-se que existem ferramentas capazes de detectar tais conteúdos nas empresas, pois as mesmas são capazes de rastrear com alta precisão infrações de direitos autorais e os métodos utilizados possuem algum grau de similaridade com os métodos necessários para atuar contra discurso de ódio.

Figura 2 – Porcentagem da citação de Nordeste por mês



Fonte: autoria própria

Legenda: Observa-se o aumento da citação especialmente nos meses de outubro e novembro. No caso do mês de outubro, a porcentagem de citações triplica em relação à média dos outros meses.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

4.2 Bag of Words

A segunda análise foi feita com a criação das BoWs dos *tweets* que possuem “nord” ou “nordestin”. Com isso, foi possível obter uma análise superficial de palavras que geralmente acompanham os *tweets* que citavam a região ou seus habitantes a cada mês. Pelo fato de a coleta ter começado na metade do mês de julho e terminado na metade do mês de dezembro, pela diferença natural na quantidade de dias de cada mês e pela quantidade de dados coletados por mês, a BoW de cada mês foi dividida pelo número total de *tweets* coletados no respectivo mês, obtendo-se, assim, a proporção de *tweets* que utilizavam cada palavra em relação aos *tweets* totais.

Foram, então, procurados adjetivos positivos ou negativos entre as palavras mais frequentes e essas palavras foram analisadas individualmente nessa etapa. Os resultados do *Bag of Words* mostram que a maior parte das palavras que estão nos *tweets* que citam o Nordeste não possui, propriamente, um significado positivo ou negativo, sendo muitas delas associadas à política ou a características da região. Esta descoberta é uma evidência que fortalece nossa justificativa de pesquisa para não ter utilizado palavras-chave muito específicas que poderiam ser usadas em discursos xenofóbicos contra nordestinos. Desta forma, podemos afirmar que a decisão foi acertada e reduziu a possibilidade de enviesar os resultados se fossem usadas palavras-chave que já carregam sentido negativo ao invés de se utilizar um adjetivo neutro que se refere à região geográfica habitada. Exemplos são: Lula, Bolsonaro, voto e semiárido. Assim, conforme observado anteriormente, foram procuradas, dentre as palavras “mineradas”, aquelas que carregam significados negativos ou positivos, dentre as quais se destacaram: “pobr” e “humild”.

4.3 Frequência de Distância

A terceira análise realizada foi a partir da criação de um modelo baseado em frequência de distância. O modelo foi criado com os dados coletados e, em seguida, foi criada uma matriz de frequência de termos de cada mês, que indica a associação entre cada uma das palavras dos textos. Por fim, foram analisadas as palavras mais associadas ao termo “nordestin” em cada mês e as diferenças nessas associações ao longo do tempo.

O modelo baseado em frequência de distância também apresentou resultados significativos. Esse modelo classifica a associação entre quaisquer duas palavras em um intervalo de -1 e 1, sendo -1 para palavras com significado semântico totalmente oposto e 1 para palavras iguais ou sinônimas.

Nos meses de julho e agosto, apenas palavras neutras aparecem entre as dez palavras mais associadas aos nordestinos. Porém, no mês de setembro, próximo às eleições, o termo “pobr” aparece na terceira posição, com 13% de associação positiva. Em outubro, “pobr” continua entre as dez mais

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

associadas, e o termo “analfabet” se encontra entre as dez, com associação de aproximadamente 10%. Nenhum termo positivo entrou entre os dez primeiros.

No mês de novembro, os termos “pobr” e “analfabet” permanecem entre os dez mais associados, entretanto, o termo “burr” alcança o segundo lugar, com 10% de associação. Novamente, nenhum termo positivo está entre os dez mais associados. Por último, em dezembro, o termo “fom” aparece na quarta posição, enquanto o termo “pobr” aparece na décima-primeira. Nenhum termo positivo ou negativo aparece entre os dez mais associados.

Por fim, em dezembro, o termo “fom” aparece na quarta posição, enquanto o termo “pobr” aparece na décima primeira. Nenhum termo positivo ou negativo está entre os dez mais associados. Apesar de fornecer algumas respostas, as associações encontradas entre as palavras não foram muito fortes. Por isso, foi requerido o uso de outro tipo de modelo para analisar esses dados.

4.4 Word2Vec

A quarta e última análise foi por meio da criação de um modelo *Word2Vec* para os dados coletados a cada mês. O modelo foi construído utilizando a biblioteca *gensim* em Python, considerando apenas as palavras que aparecem no mínimo duas vezes ao longo do mês. Esse critério foi adotado com a intenção de remover sentenças como links, sequências de *emojis* e erros mais graves de ortografia, garantindo assim um modelo mais preciso. Com esse modelo, foi possível realizar uma análise mais detalhada, identificando as palavras mais associadas a diferentes termos, além de permitir a avaliação da associação entre qualquer dupla de palavras.

Foram analisadas as palavras mais associadas ao termo “nordestin” em cada mês, observando as variações no ranking dessas palavras ao longo do tempo. Caso alguma palavra relevante para a análise aparecesse entre as mais importantes em determinado mês, mas não em outros, foi verificada a associação entre essa palavra e o termo “nordestin” nos meses subsequentes.”. Dessa forma, foram destacadas as diferenças entre os modelos mensais e construídos gráficos que facilitaram a visualização dos resultados. A partir desses gráficos e dos dados numéricos obtidos, os resultados foram interpretados e validados.

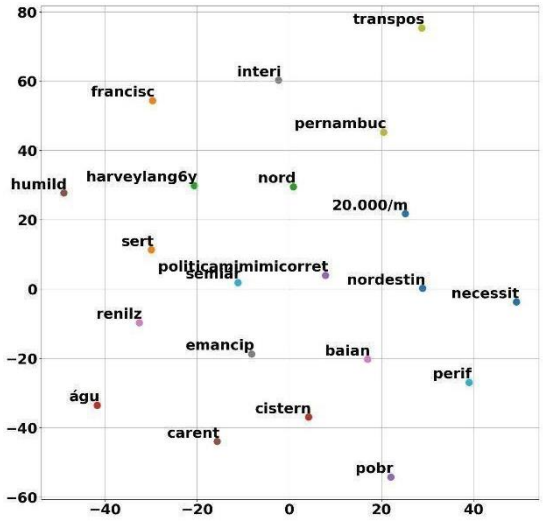
Esse modelo possibilitou a geração de um grau de associação entre todas as palavras presentes nos *tweets* em cada mês. Em geral, a maior parte dos termos mais associados a “nordestin” foram palavras neutras, como “sertão” e “água”, embora também tenham surgido alguns termos com conotação negativa.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Para interpretar os gráficos gerados por esse modelo, é importante entender que quanto mais próximo dois pontos estiverem, maior será a associação semântica entre eles. Os principais resultados de cada mês e suas respectivas representações gráficas são:

- Julho: os termos “pobr” e “humild” aparecem com aproximadamente 57% de associação ao termo “nordestin”. O termo “miser” aparece com 45%. É importante notar como o termo “nordestin” se encontra mais próximo de “interi” e “nord”, palavras neutras, conforme ilustrado na Figura 3.

Figura 3 – Gráfico representativo do modelo do mês de julho



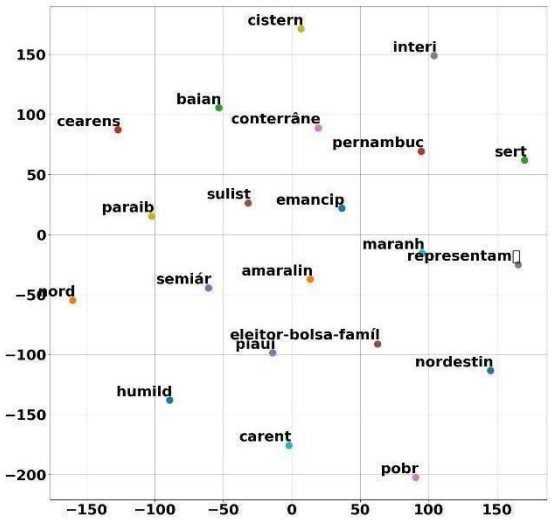
Fonte: autoria própria

Legenda: Observa-se a existência de associação entre variações da palavra “nordestino” e variações de palavras como “pobre”, “carente” e “necessitado”.

- Agosto: “pobr” aparece com 59% de associação, enquanto “humild” (mantém) 57%. “Carent” aparece com 53% e “miser” sobe para 52%. É possível notar na Figura 4 o aparecimento do termo “carent“ e como o termo “nordestin“ está mais próximo de “pobr“ do que de muitos termos neutros, como “sert“.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Figura 4 – Gráfico representativo do modelo do mês de agosto

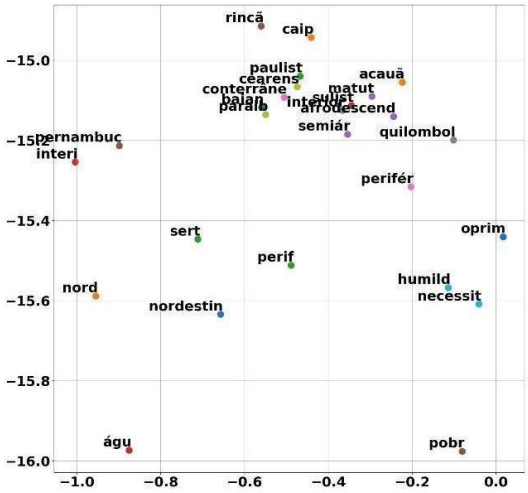


Fonte: autoria própria

- Setembro: o termo “pobr” sobe para 67% de associação, “humild” para 62%. “Carent” sobe para 54%, enquanto “miser” cai levemente para 50% (Figura 5).

Setembro: o termo “pobr” sobe para 67% de associação, “humild” para 62%. “Carent” aumenta para 54%, enquanto “miser” cai levemente para 50% (Figura 5).

Figura 5 – Gráfico representativo do modelo do mês de setembro



Fonte: autoria própria

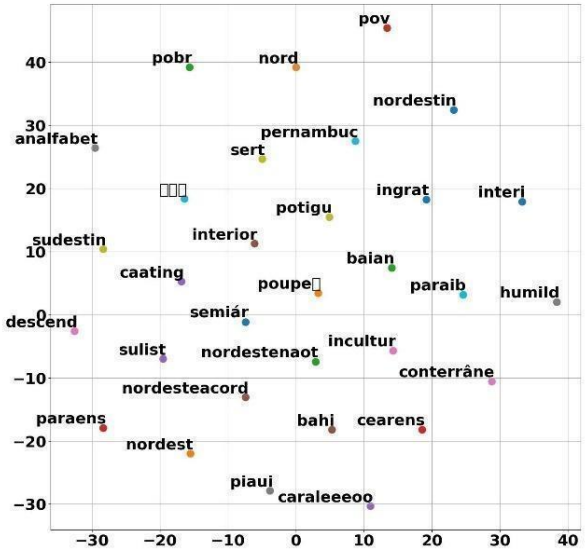
Legenda: Neste gráfico, observa-se forte associação entre variações da palavra “afrodescendente” com variações das palavras “semiárido”, “matuto”, “cearense”, “paraíba” e “quilombola”. Também

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

observa-se uma associação considerável das palavras anteriores com variações das palavras "humilde", "nordestino" e "necessitado".

- Outubro: as palavras "ingrat" e "analfabet" aparecem pela primeira vez com associação acima de 50%, respectivamente com 64% e 59% ao termo "nordestin". A palavra "pobr" mantém uma forte associação, com aproximadamente 63%, assim como "humild", que apresenta 61%. Na Figura 6, é possível observar que "pobr" está muito mais próxima do termo "nordestin" neste mês em comparação às outras palavras e ao mês de julho. Também, se nota no gráfico o surgimento dos termos "ingrat" (que se encontra mais próxima de "nordestin" do que o termo "sert" e "interi", estes últimos mais próximos em julho) e "analfabet" entre as palavras mais associadas.

Figura 6 – Gráfico representativo do modelo do mês de outubro



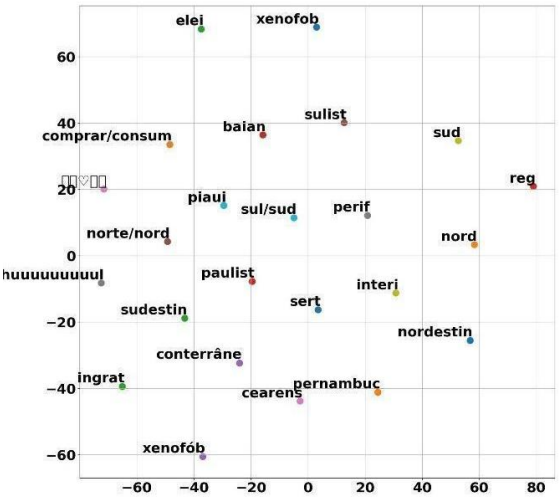
Fonte: autoria própria

Legenda: Neste caso, ocorrem associações da palavra "nordestino" com nomes de estados, mas surgem novamente variações das palavras "humilde", "pobre" e "analfabeto".

- Novembro: a palavra "ingrat" mantém associação acima de 50%, com 58%. As palavras "pobr" e "humild" caem para 57%. Na Figura 7, é possível observar que as palavras mais próximas ao termo "nordestin" são novamente palavras neutras".

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Figura 7 – Gráfico representativo do modelo do mês de novembro

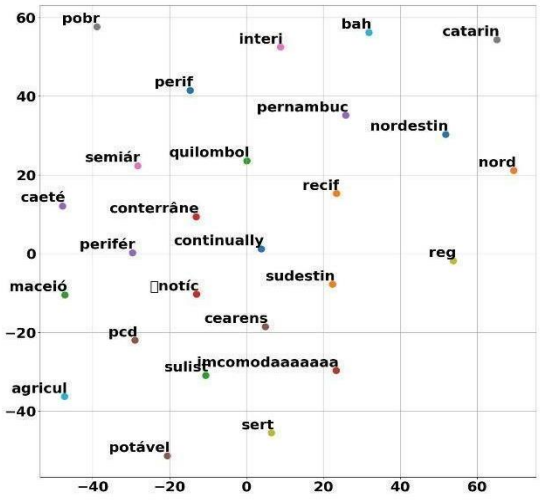


Fonte: autoria própria

Legenda: Novamente, como na Figura 6, nomes de estados ocorrem. Mas desta vez surgem variações das palavras “xenofobia” e “eleições”.

- Dezembro: a palavra “pobr” cai novamente para 54%, enquanto todos os outros termos negativos ficam abaixo de 50%. Observa-se, na Figura 8, uma maior distância do termo “pobr” e o desaparecimento dos demais termos pejorativos.

Figura 8 – Gráfico representativo do modelo do mês de dezembro



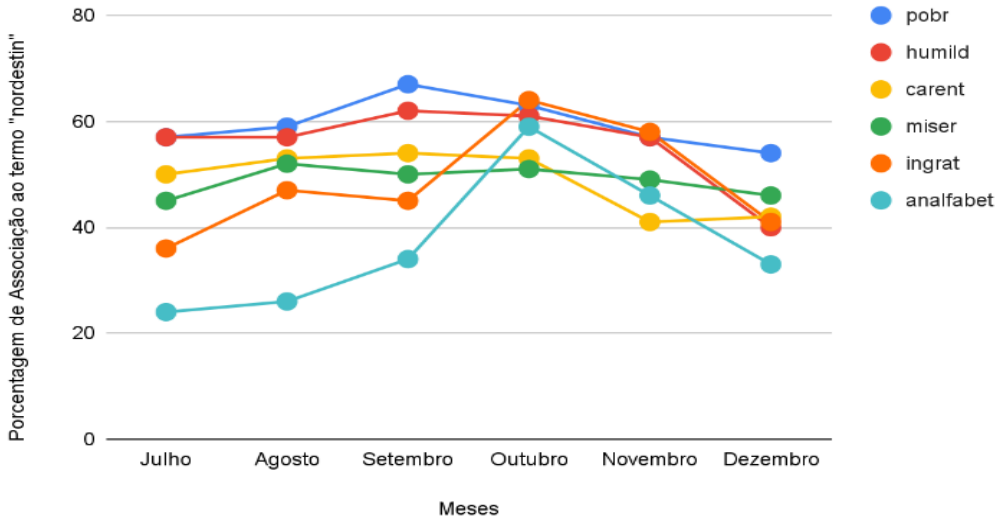
Fonte: autoria própria

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Legenda: Este gráfico, evidencia o surgimento de menções de nomes de cidades como “recife” e “maceió”, mas ainda aparecem palavras que podem estar associadas a estereótipos como “potável”, “agricultura” e “quilombola”.

Para facilitar a visualização dos dados, foi criado um gráfico (Figura 9) que apresenta as associações das palavras citadas ao longo dos meses. Nota-se que os meses das eleições (setembro e outubro) são aqueles em que os termos pejorativos atingem o seu ápice. Um fator importante a ser considerado nesta pesquisa é que as palavras-chave como “pobre”, “miserável”, “carente”, “analfabeto” não são palavras que foram escolhidas como termos de busca ou filtragem pelos autores deste artigo, mas sim palavras que surgiram dos dados a partir das análises empregadas como estando associadas às variações da palavra nordestino. Isso é uma evidência de que os textos coletados sob as condições desta pesquisa contém sentenças que associam nordestinos e o Nordeste brasileiro às ideias associadas com tais palavras-chave. Como a coleta dos dados ocorreu no período das eleições presidenciais de 2022 utilizando filtros associados às eleições, isso é uma evidência de que os discursos contidos em tais textos estavam no contexto político de xenofobia a nordestinos.

Figura 9 – Gráfico dos principais termos associados a “nordestin” que carregam conotação negativa



Fonte: autoria própria

Legenda: Deve-se notar que algumas dessas palavras apareceram nos gráficos Word2Vec e naqueles gráficos é possível notar com quais outros termos essas palavras estão associadas.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Tal representação do Nordeste em períodos eleitorais acaba por reforçar estereótipos regionais e imaginários sociodiscursivos negativos e reducionistas, evidenciando o que o historiador Durval Muniz de Albuquerque Jr. (2011, p. 33) destaca, que "o Nordeste nasce onde se encontram poder e linguagem".

Isto porque historicamente se verifica uma polarização entre o "Norte", marcado pela mestiçagem (considerada pelo darwinismo social como degeneração racial) e pelo clima árido (especialmente após a grande seca de 1877), e o "Sul", de clima mais ameno e com população de origem hegemonicamente européia e branca - a imagem do colonizador (Albuquerque Jr., 2011; Serrão, 2022; Morais, 2014). Consequentemente, o Nordeste é associado ao imaginário da seca, da pobreza, da carência e da miserabilidade, termos que encontramos em nossa pesquisa.

Além dos aspectos climáticos e populacionais, há que se destacar que, devido à crise econômica advinda do término do ciclo do açúcar, foi reforçada a percepção de desvalorização das províncias do Norte, se comparadas às do Sul do país, estimulando uma visão de que o Norte e o Nordeste eram regiões marcadas pelo atraso e decadência, em contraste ao Sul/Sudeste, avaliados como desenvolvidos.

O fim da escravidão, a expansão cafeeira e o crescente processo de industrialização e urbanização do Sul, estimularam fluxos migratórios. Enquanto a migração estrangeira era incentivada - inclusive custeada - pelo Estado brasileiro, os migrantes oriundos do Norte/Nordeste recebiam tratamento extremamente desigual, aos quais competiam as tarefas mais árduas e cuja precariedade de condições assemelhava-se a um regime de servidão, de acordo com Celso Furtado (2007, p. 196), quando analisa a formação econômica do Brasil. A migração em massa na década de 1930 do Nordeste rural para o Sudeste em crescente industrialização contribuiu, ainda, para o aumento do preconceito contra os nordestinos (Serrão, 2022).

Os dados deste estudo demonstram que tais estereótipos históricos e imaginários sociodiscursivos a respeito do Nordeste ancoram o discurso xenófobo e de ódio que têm sido acionado em períodos eleitorais para desqualificação do eleitorado e de candidatos oriundos do Nordeste, fenômeno muitas vezes travestido de "liberdade de expressão".

Internacionalmente, a ONU - Organização das Nações Unidas prevê responsabilidades no exercício da liberdade de expressão e restrições em favor da guerra e a apologia do ódio nacional, radical, racial ou religioso que constitua incitamento à discriminação, à hostilidade ou à violência – o discurso de ódio (Silva, 2016). No Brasil, especificamente, a estes casos tem-se aplicado

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

timidamente a legislação relativa à discriminação e intolerância, especialmente a Lei n. 7.716/1989, conhecida como Lei Antirracismo.

Considerando que a violação à dignidade presenciada via discurso odioso e xenófobo transcende a esfera dos direitos individuais, a mesma vulnera a democracia e a construção de uma sociedade justa, fundada na igualdade e na diversidade. Por tal motivo que estudos como este buscam dar visibilidade ao problema social do discurso de ódio e, em específico, do discurso xenófobo.

5. CONCLUSÃO

Os resultados desta pesquisa confirmam que as eleições presidenciais de 2022 exerceram um impacto significativo da visão negativa em relação ao Nordeste e ao povo nordestino, evidenciando que esse período se configurou como um momento crítico para o crescimento do discurso de ódio. Conforme afirma Jenkins (2009), a política do ataque evolui no nível popular, e cada vez mais as pessoas manifestam desacordo por meio da difamação daqueles que realizam escolhas políticas distintas das suas.

Os resultados obtidos pelos modelos demonstraram que, à medida que os *tweets* se aproximavam dos meses das eleições (setembro e outubro), as palavras mais comumente associadas ao Nordeste e aos nordestinos se tornaram progressivamente mais negativas, atingindo seu ápice no mês do pleito.

Observa-se também que nenhuma palavra com conotação positiva está entre as mais associadas ao termo "nordestino", sendo todas neutras ou negativas. Portanto, conclui-se que a visão dos usuários do Twitter que comentaram sobre as eleições é, em maior parte, pejorativa, com a negatividade aumentando conforme a proximidade do pleito.

Sob uma perspectiva maniqueísta e dicotômica, os nordestinos são representados como uma ameaça à nação devido ao seu comportamento eleitoral. Termos como "pobre", "miserável", "analfabeto", "burro", "ingrato" associam os nordestinos ao atraso, a severas limitações cognitivas, ao analfabetismo, à pobreza e à dependência de políticas "assistencialistas". Exemplos emblemáticos *tweets* como: *"Esse povo do Nordeste burro não sabe votar", "Nordeste tem que se fuder como sempre, povo burro"*, "*@jairbolsonaro se ganhar corta a água (que já é limitada) e corta fora do Bolsa Família, e cortá-los da porra do mapa e deixá-los ser um lugar independente para que o comunismo possa ficar lá, pelo amor de Deus".*

Tais discursos são frequentemente utilizados para desqualificar e desumanizar esta importante parcela da população brasileira, inserindo-se no espectro da xenofobia. De forma similar ao estudo

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

de Iasulaitis *et al.* (2023), foi possível identificar o uso intenso das redes sociais para a realização de campanha negativa.

Com este trabalho, pretende-se contribuir para melhor compreensão do fenômeno do discurso de ódio com teor xenófobo no Brasil, possibilitando a promoção de ações de prevenção, formação cidadã, defesa dos direitos humanos, bem como o desenvolvimento de políticas públicas voltadas ao combate às injustiças sociais, visando a promoção de um Brasil unido em sua diversidade.

REFERÊNCIAS

ALBUQUERQUE JR., Durval Muniz **A invenção do Nordeste e outras artes**. Prefácio de Margareth Rago. São Paulo: Cortez, 2011.

BARBERÁ, Pablo; RIVERO, Gonzalo. Understanding the political representativeness of Twitter users. **Social Science Computer Review**, v. 33, n. 6, p. 712-729, 2015.

BARBOSA, Iann Carvalho et al. Reconhecimento de mensagens com teor transfóbico no Twitter. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Federal de Campina Grande, Campina Grande, 2021.

BOURDIEU, Pierre. **O poder simbólico**. 10. ed. Rio de Janeiro: Bertrand Brasil, 2007.

BOYD, Danah M.; ELLISON, Nicole B. Social network sites: Definition, history, and scholarship. **Journal of computer-mediated Communication**, v. 13, n. 1, p. 210-230, 2007.

BRUGGER, Winfried. Proibição ou proteção do discurso do ódio? Algumas observações sobre o direito alemão e o americano. **Direito Público**, v. 4, n. 15, 2007.

DE ALBUQUERQUE JÚNIOR, Durval Muniz. **A invenção do Nordeste e outras artes**. Cortez editora, 2021.

DE ARAUJO, Gabriela Denise; DE MORAES, Fabricio Landi; PISA, Ivan Torres. Análise Exploratória de dados do Twitter: compreendendo as conexões da informação de saúde durante o surto da febre amarela em 2017. **Brazilian Journal of Information Science**, v. 14, n. 3, p. 6, 2020.

FURTADO, Celso. Formação econômica do Brasil. São Paulo: Companhia das Letras, 2007.

HOEPFNER, Soraya Guimarães. Apontamentos sobre a questão ético-midiática do discurso de ódio na rede social. **Esferas**, n. 4, 2014.

IASULAITIS, Sylvia; VICARI, Isabella. The Salience of Traditional Moral Values: Bolsonaro's Electoral Competition Strategy on Twitter. **Int'l J. Soc. Sci. Stud.**, v. 9, p. 153, 2021.

IASULAITIS, Sylvia; VALEJO, Alan Demétrius Baria; GRECO, Bruno Cardoso; PERILLO, Vinicius Gonçalves; MESSIAS, Guilherme Henrique; VICARI, Isabela. The Interfaces Twitter Elections Dataset: Construction process and characteristics of big social data during the 2022 presidential elections in Brazil. **PLoS ONE** 20(2): e0316626. https://doi.org/10.1371/journal.pone.0316626

IASULAITIS, Sylvia; VIEIRA, Aiane Oliveira; SELEGHIM, Ariane Duarte. CAMPANHAS ELEITORAIS MULTIPLATAFORMAS EM TEMPOS DE CONVERGÊNCIA MIDIÁTICA. **Revista GEMInIS**, v. 14, n. 1, p. 121-148, 2023.

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

JENKINS, Henry. Cultura da convergência. Tradução: Suzana Alexandria. São Paulo. **Aleph**, v. 2, 2009.

KANNAN, Subbu *et al*. Preprocessing techniques for text mining. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7-16, 2014.

LIDDY, Elizabeth D. **Natural language processing**. School of Information Studies - Faculty Scholarship, 2001.

MARTINEZ, Wendy L.; MARTINEZ, Angel R.; SOLKA, Jeffrey. **Exploratory data analysis with MATLAB**. Chapman and Hall/CRC, 2017.

MARTINS, Anna Clara Lehmann. Discurso de ódio em redes sociais e reconhecimento do outro: o caso M. **Revista Direito GV**, v. 15, p. e1905, 2019.

MORAIS, Argus Romero Abreu de. Os imaginários sociodiscursivos acerca do nordeste brasileiro. **EID&A - Revista Eletrônica de Estudos Integrados em Discurso e Argumentação**, Ilhéus, n. 7, p. 22-38, dez.2014.

MULLAH, Nanlir Sallau; ZAINON, Wan Mohd Nazmee Wan. Advances in machine learning algorithms for hate speech detection in social media: a review. **IEEE Access**, v. 9, p. 88364-88376, 2021.

NOCKLEBY, John T. et al. Encyclopedia of the American constitution. **Detroit, MI: Macmillan Reference**, v. 3, n. 2, 2000.

ORIOLA, Oluwafemi; KOTZÉ, Eduan. Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. **IEEE Access**, v. 8, p. 21496-21509, 2020.

SERRAO, Rodrigo. Racializing region: Internal Orientalism, social media, and the perpetuation of stereotypes and prejudice against Brazilian Nordestinos. **Latin American Perspectives**, v. 49, n. 5, p. 181-199, 2022.

SILVA, Rosane Leal da et al. Discursos de ódio em redes sociais: jurisprudência brasileira. **Revista direito GV**, v. 7, p. 445-468, 2011.

SILVA, Yane M. P. "Esses Nordestinos… Discurso de Ódio em Redes Sociais da Internet na Eleição Presidencial de 2014. Brasília/DF, setembro de 2016.

Informações sobre o Artigo

Fonte de financiamento: Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, processo n. 2022/03090-0.

Gustavo Gurjão Camargo Campos

Bacharel em Ciência da Computação formado pela Universidade Federal de Campina Grande (UFCG). Durante sua formação, integrou as atividades e ações do Grupo de Pesquisa Interfaces - Núcleo de Estudos Sociopolíticos dos Algoritmos e da Inteligência Artificial.

E-mail: gustavo.campos@ccc.ufcg.edu.br

ORCID: https://orcid.org/0009-0009-2651-6682

Revista GEMInIS
ISSN: 2179-1465
https://www.revistageminis.ufscar.br

Eanes Torres Pereira

Professor da Unidade Acadêmica de Sistemas e Computação da Universidade Federal de Campina Grande. Pesquisador em aplicações de inteligência artificial para reconhecimento de padrões em dados multimídia. Atua nos seguintes grupos de pesquisa: Grupo de Pesquisa Interfaces - Núcleo de Estudos Sociopolíticos dos Algoritmos e da Inteligência Artificial; Grupo de Inteligência Artificial, Ciência de Dados e Arquiteturas Embarcadas.

E-mail: eanes@computacao.ufcg.edu.br

ORCID: https://orcid.org/0000-0002-9717-794X

Sylvia Iasulaitis

Professora Doutora da Universidade Federal de São Carlos (Brasil) e Honorary Research Fellow da Liverpool Hope University (Inglaterra). Docente permanente dos Programas de Pós-Graduação em Ciência, Tecnologia e Sociedade e de Ciência da Informação da UFSCar. Lidera o Interfaces - Núcleo de Estudos Sociopolíticos dos Algoritmos e da Inteligência Artificial. Atua nas áreas de Ciência Social Computacional e Ciência de Dados Sociais.

E-mail: si@ufscar.br

ORCID: https://orcid.org/0000-0002-3526-1003