
**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE**
**CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA**
**CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**GUSTAVO GURJÃO CAMARGO CAMPOS**

# ANÁLISE TEXTUAL DE TWEETS REFERENCIANDO NORDESTINOS DURANTE AS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS DE 2022

**CAMPINA GRANDE - PB**

**2023**

**GUSTAVO GURJÃO CAMARGO CAMPOS**

# ANÁLISE TEXTUAL DE TWEETS REFERENCIANDO NORDESTINOS DURANTE AS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS DE 2022

**Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.**

**Orientador: Professor Dr. Eanes Torres Pereira**

**CAMPINA GRANDE - PB**

**2023**

**GUSTAVO GURJÃO CAMARGO CAMPOS**

# ANÁLISE TEXTUAL DE TWEETS REFERENCIANDO NORDESTINOS DURANTE AS ELEIÇÕES PRESIDENCIAIS BRASILEIRAS DE 2022

**Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.**

## BANCA EXAMINADORA:

**Professor Dr. Eanes Torres Pereira**

**Orientador – UASC/CEEI/UFCG**

**Professor Dr. Wilkerson de Lucena Andrade**

**Examinador – UASC/CEEI/UFCG**

**Professor Tiago Lima Massoni**

**Professor da Disciplina TCC – UASC/CEEI/UFCG**

**Trabalho aprovado em: 14 de Fevereiro de 2023.**

**CAMPINA GRANDE - PB**

# ABSTRACT

With the invention of social networks and their popularization, they became platforms of discussion about various subjects, including politics. During the 2022 Brazilian presidential election, that was also the case. The Northeast of Brazil, being a highlight of the 2022 presidential elections, was also the subject of discussion during this period. So, this paper has the intention of analyzing mentions of people from the Northeast, during the pre-elections and the elections. This project used data preprocessing, by making use of natural language processing methods and the analysis was made using exploratory textual analysis techniques and algorithms, such as Word2Vec and the concept of frequency of distance. Evidences that support the idea that during the election period, the closer it was from the voting day, the more negative the opinion of Twitter users about people from the Northeast was were found.

# Análise textual de tweets referenciando nordestinos durante as eleições presidenciais brasileiras de 2022

Gustavo Gurjão Camargo Campos
gustavo.campos@ccc.ufcg.edu.br
Universidade Federal de Campina Grande

Eanes Torres Pereira
eanes@computacao.ufcg.edu.br
Universidade Federal de Campina Grande

## Resumo

Com o surgimento das redes sociais e sua popularização, elas se tornaram plataformas de discussão sobre os mais variados assuntos, incluindo a política. No cenário das eleições presidenciais brasileiras de 2022, não foi diferente. O Nordeste, tendo destaque durante as eleições de 2022, também foi assunto de discussões durante esse período, então esse projeto tem o objetivo analisar as menções a nordestinos em *tweets* durante o período de pré eleições e de eleições. Esse projeto empregou pré-processamento de dados, utilizando métodos de processamento de linguagem natural e a análise dos dados foi feita por meio de técnicas e de algoritmos de análise exploratória textual, como o Word2Vec e a frequência de distância. Foram encontradas evidências de que no decorrer do período das eleições, quanto mais se aproximava os dias de votação, mais negativa era a opinião dos usuários da rede social acerca dos nordestinos.

## Abstract

With the invention of social networks and their popularization, they became platforms of discussion about various subjects, including politics. During the 2022 Brazilian presidential election, that was also the case. The Northeast of Brazil, being a highlight of the 2022 presidential elections, was also the subject of discussion during this period. So, this paper has the intention of analyzing mentions of people from the Northeast, during the pre-elections and the elections. This project used data preprocessing, by making use of natural language processing methods and the analysis was made using exploratory textual analysis techniques and algorithms, such as Word2Vec and the concept of frequency of distance. Evidences that support the idea that during the election period, the closer it was from the voting day, more negative the opinion of Twitter users about people from the Northeast was were found.

## Keywords

Nordeste, Word2vec, Twitter, Eleições Presidenciais 2022

## 1 INTRODUÇÃO

Com o desenvolvimento das tecnologias e os avanços dos meios de comunicação, as redes sociais se tornaram o centro das interações entre as pessoas no meio digital, por conta de sua acessibilidade e facilidade de uso. Como um reflexo da sociedade, as redes sociais são, hoje, um espaço de discussões acerca de diferentes tópicos, como culturais, de entretenimento, sociológicos, científicos e políticos [1].

Entre os tópicos políticos discutidos no ano de 2022, as eleições presidenciais, as mais importantes do país, foram um dos maiores destaques. Com a realização de debates, a reserva de horários voltados para propagandas eleitorais e a discussão do tópico em jornais na televisão aberta, fica claro o quanto as eleições presidenciais tomaram conta das notícias do ano.

Os resultados das eleições presidenciais determinaram a vitória do candidato Luís Inácio "Lula" da Silva sobre o então presidente Jair Messias Bolsonaro. A nível estadual, Lula garantiu vitória em treze dos vinte e seis estados brasileiros, sendo nove deles: Rio Grande do Norte, Paraíba, Pernambuco, Sergipe, Aracaju, Ceará, Bahia, Maranhão e Piauí, os nove estados do Nordeste. Por causa dessa situação, o Nordeste surgiu como grande responsável pelo resultado das eleições no meio popular, garantindo uma grande variedade de comentários acerca da região e de seus habitantes nas redes sociais em relação ao tópico das eleições [2].

O Twitter é uma das redes sociais mais utilizadas no Brasil [3]. Conhecida por sua política de limitação de 280 caracteres por postagem desde 2017, é um meio pelo qual as pessoas podem fácil 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 uma rede social propícia para análises textuais.

Este trabalho procura investigar, através da coleta de postagens do Twitter, a variação das opiniões dos usuários dessa rede social acerca do Nordeste. Analisando a evolução do destaque da região na rede social à medida que os meses de votação se aproximavam e a opinião de seus usuários sobre os nordestinos durante esse período, para checar se essa opinião se tornou mais negativa ou positiva à medida que essa aproximação aumentava. Com isso, seria possível identificação automática de possíveis discursos de ódio e outros crimes virtuais ligados à xenofobia contra o povo nordestino. Os métodos escolhidos procuram buscar a associação do termo "nordestino" e suas variações a diferentes palavras e como essa associação muda com o passar dos meses, especialmente focando se palavras pejorativas têm sua associação aumentada, diminuída ou mantida estável quanto mais se aproximavam os meses do pleito.

## 2 TRABALHOS RELACIONADOS

Mullah e Zainon [4], em seu artigo, citam que a evolução das técnicas de aprendizagem de máquina para classificação e para detecção de discurso de ódio somada ao avanço no processamento da linguagem natural contribuiu para a realização de estudos em diferentes países, contextos e redes sociais usando aprendizagem de máquina, fornecendo resultados interessantes.

Já foram realizados estudos de detecção de linguagem de ódio no Twitter, mais precisamente no contexto da África do Sul. Nesse estudo, Oriola e Kotzé [5] usaram técnicas de Aprendizagem de Máquina e de classificação para detectar e classificar *tweets* que continham esse tipo de conteúdo. O diferencial desse trabalho se deve ao fato de não ter ocorrido nenhum estudo prévio que focasse em seu país, o que motivou os autores a prosseguir com a pesquisa.

Eles obtiveram sucesso com modelos baseados em *Support Vector Machine* (Máquina de Vetores de Suporte) e *Random Forest* (Floresta Aleatória), porém concluíram que seria interessante a tentativa de utilização de modelos híbridos que poderiam levar a melhores resultados.
Membros da Universidade Federal de São Paulo (Unifesp) [6] realizaram estudos no Twitter envolvendo análise exploratória de dados, analisando textos e hashtags, mostrando que é possível a realização de estudos de análise descritiva nessa plataforma. Outro projeto que utilizou o Twitter foi o de Iaan Carvalho Barbosa da UFCG [7], que utilizou dados coletados da rede social para construir uma rede neural que detectava linguagem transfóbica.

# 3 FUNDAMENTAÇÃO TEÓRICA

Esta seção explana acerca do Twitter, processamento de linguagem natural e conceitos de análise exploratória de dados textuais e mineração textual.

## 3.1 REDES SOCIAIS E TWITTER

Redes sociais são um conjunto constituído por atores, que são as pessoas que as utilizam e as interações entre eles. Uma rede social possui atores, postagens, perfis, sistemas de avaliação (como por exemplo, os *"likes"*) e cada rede social possui diferentes características que as distinguem umas das outras [8]. Para o desenvolvimento deste trabalho, a rede social escolhida foi o Twitter, por apresentar um limite de caracteres em suas postagens, (chamadas de *"tweets"*), facilitando a coleta e análise textual das mesmas.

## 3.2 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial que tem o objetivo de fazer os computadores entenderem, interpretarem e manipularem a linguagem humana. Um dos conceitos dentro da área de PLN é o de pré-processamento textual. O pré-processamento textual se refere ao processo de transformação de um texto do estado de sua coleta, para um estado em que ele possa ser analisado na realização da tarefa alvo [9]. Neste trabalho, foram realizadas algumas técnicas necessárias para a realização do pré-processamento dos dados [10]:

- `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".
- `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".

## 3.3 ANÁLISE EXPLORATÓRIA DE DADOS

A análise exploratória de dados textuais se refere ao conjunto de técnicas usadas para analisar e investigar conjunto de dados, resumir suas principais características, investigar padrões encontrados, testar hipóteses e facilitar sua visualização e leitura [11]. As técnicas e algoritmos usados para analisar os dados nesse projeto, foram:

- `Bag of Words`: representação de um conjunto de dados de acordo com a ocorrência de cada palavra ou *token* encontrado nele.
- `Frequência de Distância`: técnica que mapeia *tokens* com sua frequência num texto. Com ela, também é possível prever uma associação entre termos ou *tokens* presentes em um conjunto de dados baseada em suas respectivas frequências.
- `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.
- `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). Por exemplo, na Figura 1, o termo "nordestin" é mais semanticamente próximo ao termo "sert" do que ao termo "agu".

# 4 METODOLOGIA

Para que essa pesquisa fosse concluída, foram seguidos os passos: coleta de dados, pré-processamento desses dados e a realização de quatro diferentes análises com diferentes métodos com a intenção de descobrir a mudança da opinião dos usuários do Twitter acerca do povo nordestino no decorrer do período de eleições.

## 4.1 COLETA DOS DADOS

Para a coleta dos dados, foram criadas contas no *Developer Twitter*[1], 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 república (Lula e Bolsonaro). Esses dados são uma amostra da coleta feita por um grupo de pesquisadores compostos por alunos e professores da Universidade Federal de Campina Grande e da Universidade Federal de São Carlos e fazem parte de

[1]Encontrado no link: https://developer.twitter.com/en. Acesso: 25/01/2023, 11:00

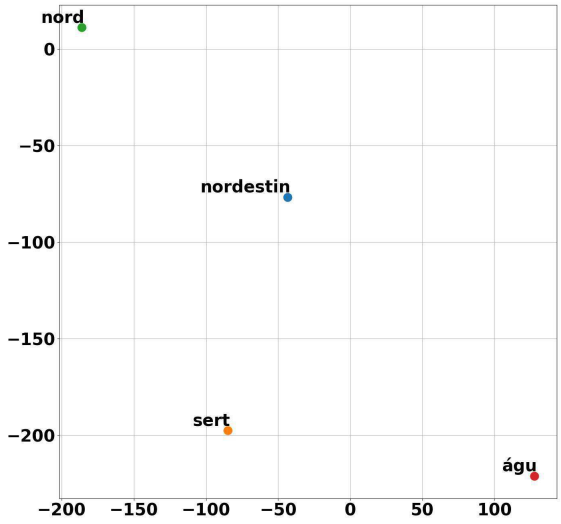


**Figura 1: Exemplo de representação gráfica gerada a partir de um modelo Word2Vec**

um projeto maior que procura realizar diferentes análises acerca das eleições, sendo esta pesquisa uma delas.

## 4.2 PRÉ-PROCESSAMENTO

Com os dados coletados, a primeira ação realizada foi o pré-processamento dos dados textuais. Os *tweets* recebidos passaram primeiramente pelo processo de lowercasing, deixando todas as letras minúsculas. Depois, passaram para a etapa de tokenização, transformando cada *tweet* em listas das palavras contidas neles. Com essas listas prontas, o processo seguinte foi o de stemização, reduzindo cada palavra à sua forma raiz. Por último, foram retirados os caracteres especiais e as *stopwords*, com a intenção de maximizar a possibilidade de extrair apenas significados semânticos relevantes.

## 4.3 PRESENÇA NOS *TWEETS*

Primeiramente, foi analisada a presença de palavras e termos se referindo 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 possuem as formas raiz "nord" (forma-raiz de "nordeste") e "nordestin" (forma raiz de nordestino, nordestina, nordestinos, nordestinas e afins) em relação aos *tweets* totais. Assim, é possível ver em que meses as palavras relacionadas aos nordestinos tiveram mais impacto nos *tweets* e que impacto foi esse em comparação aos outros meses.

## 4.4 BAG OF WORDS

A segunda análise feita foi a criação das *Bag of Words* (BoW) dos *tweets* que possuem "nord" ou "nordestin". Com isso, é possível ter uma análise superficial de palavras que geralmente acompanham os *tweets* que citam 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 da 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 mesmo, tendo assim a proporção de *tweets* que utilizam 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.

## 4.5 FREQUÊNCIA DE DISTÂNCIA

A terceira análise realizada foi a criação de um modelo baseado em frequência de distância. O modelo foi criado com os dados coletados e então, foi criada uma matriz de frequência de termos de cada mês que indica a associação entre cada uma das palavras dos textos. Por último, foram analisadas as palavras mais associadas ao termo "nordestin" em cada mês e as diferenças dessas palavras com o passar dos meses.

## 4.6 WORD2VEC

A quarta e última análise foi a criação de um modelo Word2Vec com os dados recebidos para cada mês. Foi criado um modelo a partir da biblioteca gensim[1] em Python, com a inserção de palavras que aparecem no mínimo duas vezes ao longo do mês, com a intenção de remover palavras, como links, sequências de emojis e erros mais graves de ortografia, garantindo, assim, um modelo mais preciso. Com esse modelo, foi feita uma análise mais detalhada, sendo possível ver as palavras mais associadas a diferentes termos, além de ser possível exibir a associação entre qualquer dupla de palavras. Foram analisadas as palavras mais associadas ao termos "nordestin" em cada mês, observando-se as diferentes palavras que aparecem neste ranking com o passar do tempo. Além disso, se alguma palavra relevante para análise aparecesse entre as mais importantes de um mês, mas não em outro, também foi checada a associação entre essas palavras e o termo "nordestin" nesses meses. Foram então, destacadas as diferenças entre cada modelo de cada mês e construídos gráficos que auxiliam na visualização dos resultados. Com os gráficos feitos e os dados numéricos encontrados, os resultados foram interpretados e validados, ao observar as diferenças entre os modelos ao longo do tempo.

# 5 RESULTADOS

Nesta seção, serão apresentados os resultados de cada etapa do projeto.

## 5.1 COLETA DE DADOS

Os resultados da coleta de dados resultaram em uma planilha com as colunas (Figura 2):

- `id`:identificador do dado;
- `text`:o conteúdo textual do *tweet*;
- `created_at`: data e horário do *tweet*;
- `source`:plataforma de origem do *tweet*;
- `lang`: linguagem do *tweet*;

[1]Encontrada no link: https://radimrehurek.com/gensim/models/word2vec.html. Acesso: 25/01/2023, 11:00

- `conversation_id`: identificador da thread do *tweet*;
- `like_count`:quantidade de curtidas do *tweet*;
- `retweet_count`:quantidade de *retweets* do *tweet*;
- `quote_count`:quantidade de *quote tweets* do *tweet*;
- `reply_count`:quantidade de respostas ao *tweet*;
- `type`:se o *tweet* foi uma postagem original ou uma resposta a outro *tweet*;
- `referenced_tweet_id`:caso seja um quote tweet ou uma resposta a um *tweet*, id do *tweet* original;
- `mentions`: id dos *tweets* que o *tweet* referenciou;
- `hashtags`: *hashtags* utilizadas no *tweet*;
- `urls`: *links* utilizados no *tweet*;
- `author_id`: id do autor do *tweet*;

| | id | text | created_at | source | lang | conversation_id |
|---|---|---|---|---|---|---|
| 0 | 1549544611860959232 | "esmagadora " — https://t.co/nWhnD9f5sX | 2022-07-19 23:59:59 | Twitter Web App | pt | 1549544611860959232 |
| 1 | 1549544600347492353 | @talphavirginis_ @adaniellelouise Amiga, até o... | 2022-07-19 23:59:56 | Twitter Web App | pt | 1549536560525213696 |
| 2 | 1549544587336753153 | LULA LADRÃO, ROUBOU MEU CORAÇÃO | 2022-07-19 23:59:53 | Twitter for iPhone | pt | 1549544587336753153 |

**Figura 2: Primeiras colunas e linhas de um resultado de coleta de Julho**

## 5.2 PRÉ-PROCESSAMENTO

Para a pesquisa, só foi utilizada a coluna *"text"* dos dados coletados, logo, o pré-processamento foi feito apenas nessa coluna textual, com as etapas descritas na seção 4.2. O pré-processamento resultou em uma lista de palavras-raiz que não são consideradas *stopwords*, como é possível ver na Figura 3.

| | id | text | created_at | source | lang | conversation_id |
|---|---|---|---|---|---|---|
| 0 | 1549544611860959232 | [esmag, —, http, //t.co/nwhnd9f5sx] | 2022-07-19 23:59:59 | Twitter Web App | pt | 1549544611860959232 |
| 1 | 1549544600347492353 | [talphavirginis_, adaniellelouis, amig, ond, s... | 2022-07-19 23:59:56 | Twitter Web App | pt | 1549536560525213696 |
| 2 | 1549544587336753153 | [lul, ladr, roub, coraçã] | 2022-07-19 23:59:53 | Twitter for iPhone | pt | 1549544587336753153 |

**Figura 3: Resultado do pré-processamento da Figura anterior**

## 5.3 PRESENÇA NOS *TWEETS*

Como é possível ver na Figura 4 resultante, a proporção de postagens que citam o Nordeste aproximadamente triplica 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 contendo o segundo turno das eleições.

## 5.4 BAG OF WORDS

Os resultados do Bag of Words mostram que a maior parte das palavras que estão nos *tweets* que citam o Nordeste são palavras que não possuem um significado positivo ou negativo, muitas delas,

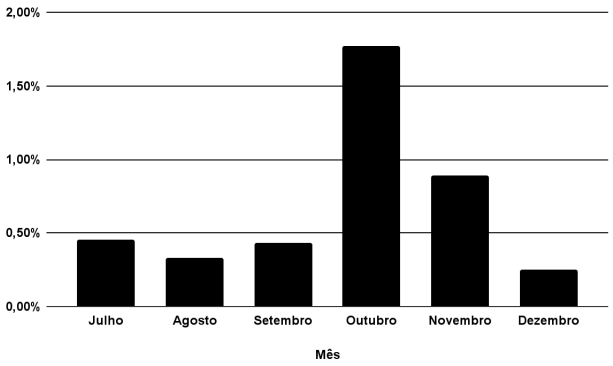


**Figura 4: Porcentagem da citação de Nordeste por mês**

associadas à política ou a características da região. Exemplos são: Lula, Bolsonaro, voto e semiárido. Então, como dito anteriormente, foram procuradas palavras que carregam significados negativos ou positivos, delas as de maior destaque foram: "pobr" e "humild".

## 5.5 FREQUÊNCIA DE DISTÂNCIA

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

Nos meses de Julho e Agosto, apenas palavras neutras aparecem nas 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. No mês de Outubro, "pobr" continua entre as dez mais associadas, mas também, 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" estão entre as dez mais associadas, porém o termo "burr" adquire o segundo lugar, com 10% de associação. Novamente, nenhum termo positivo entrou entre os dez mais associados.

Por último, no mês de 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.

Apesar de providenciar 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.

## 5.6 WORD2VEC

Na criação do modelo Word2Vec, a limitação utilizada foi a de excluir palavras que aparecem apenas uma vez nos textos. Isso foi feito para evitar que *links* e erros de digitação mais graves não influenciem nos resultados. Com esse modelo, foi possível gerar um grau de associação entre todas as palavras presentes nos *tweets* em cada mês. Novamente, a cada mês, a maior parte dos termos mais associados a nordestinos são termos neutros, como sertão e água, porém, também há alguns termos de conotação negativa. Para

interpretar um gráfico gerado por esse modelo, é apenas necessário entender que, quanto mais próximo estão dois pontos, maior 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%. Note como o termo "nordestin" se encontra mais próximo de "interi" e "nord", palavras neutras, na Figura 5.

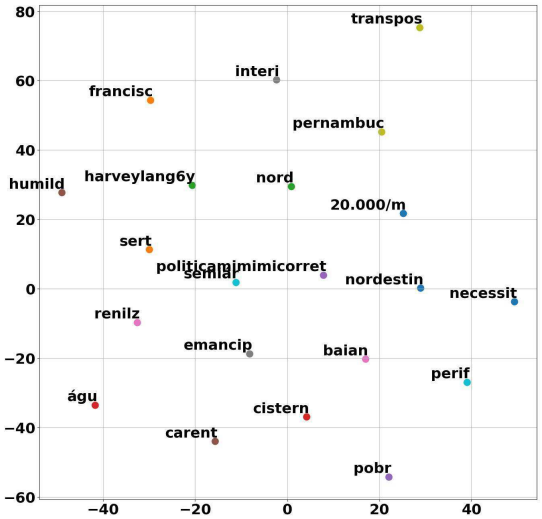


**Figura 5: Gráfico representativo do modelo do mês de Julho**

- `Agosto`: "pobr" aparece com 59% de associação, "humild" continua com 57%. "carent" aparece com 53% e "miser" subiu para 52%. É possível notar na Figura 6 o aparecimento do termo "carent" e como o termo "nordestin" está mais próximo de "pobr" do que de muitos termos neutros, como "sert".
- `Setembro`: "pobr" sobe para 67% de associação, "humild" sobe para 62%. "carent" sobe para 54% e "miser" cai levemente para 50% (Figura 7).
- `Outubro`: as palavras "ingrat" e "analfabet" aparecem pela primeira vez acima de 50% com 64% e 59% de associação à palavra "nordestin" respectivamente. A palavra "pobr" continua com forte associação com aproximadamente 63%, assim como humilde, que apresenta 61%. É possível notar como a palavra "pobr" se encontra muito mais próxima ao termo "nordestin" neste mês em relação às outras palavras em comparação ao mês de Julho. Também, nota-se no gráfico o surgimento dos termos "ingrat" (que se encontra mais próxima de "nordestin" do que o termo "sert" e "interi", que aparecia mais próximo em Julho) e "analfabet" nas palavras mais associadas (Figura 8).

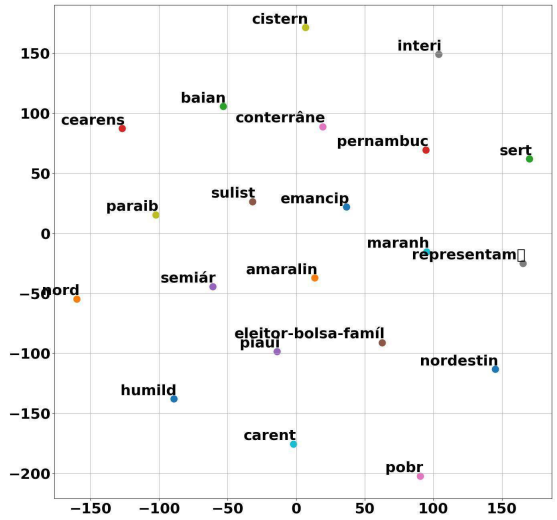


**Figura 6: Gráfico representativo do modelo do mês de Agosto**

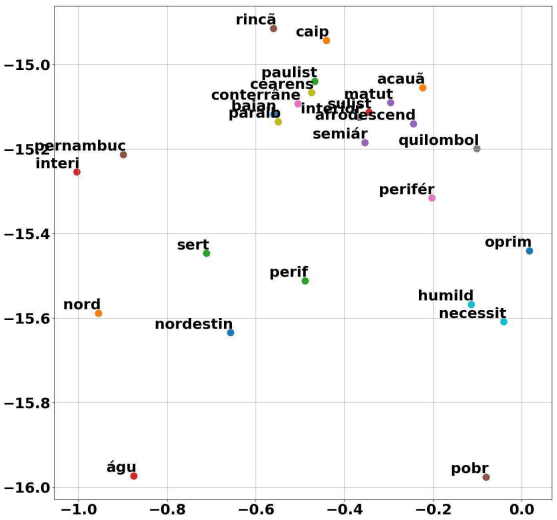


**Figura 7: Gráfico representativo do modelo do mês de Setembro**

- `Novembro`: a palavra "ingrat" continua acima de 50% com 58% de associação, as palavras "pobr" e "humild" caem para 57%. É possível ver na Figura 9 que as palavras mais próximas ao termo "nordestin" são novamente palavras neutras.

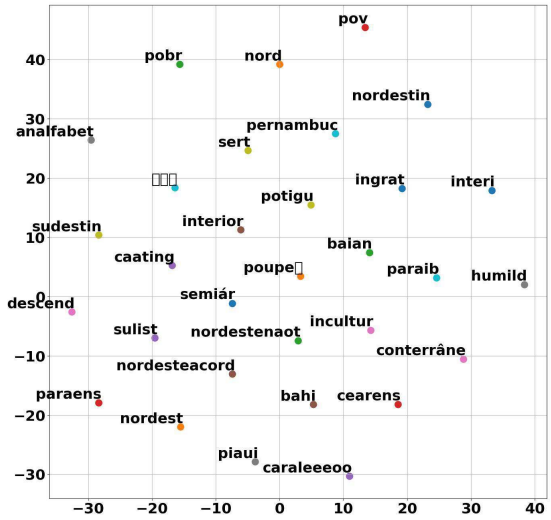


**Figura 8: Gráfico representativo do modelo do mês de Outubro**

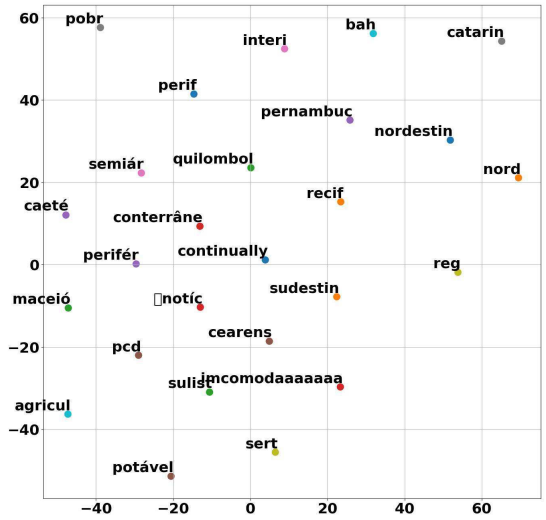


**Figura 10: Gráfico representativo do modelo do mês de Dezembro**

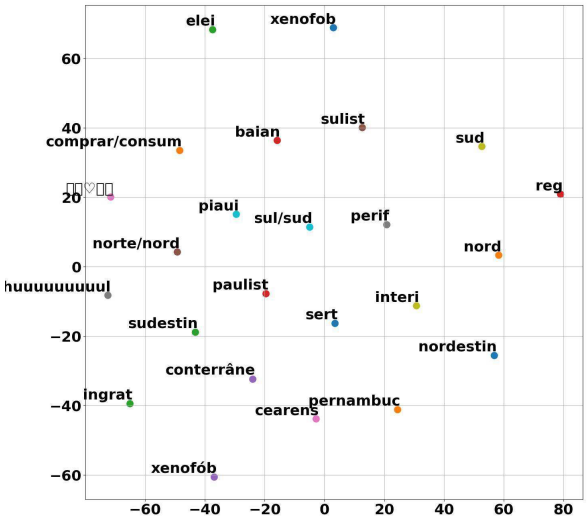


**Figura 9: Gráfico representativo do modelo do mês de Novembro**

- `Dezembro`: "pobr" cai novamente para 54% e todos os outros termos negativos ficam abaixo de 50%. Nota-se uma maior distância na figura do termo "pobr" e o desaparecimento de outros termos pejorativos (Figura 10).

Para facilitar as informações, foi criado um gráfico com as associações das palavras citadas ao longo dos meses. Nota-se que os meses das eleições (Setembro e Outubro) são os meses em que os termos pejorativos estão no seu ápice (Figura 11).

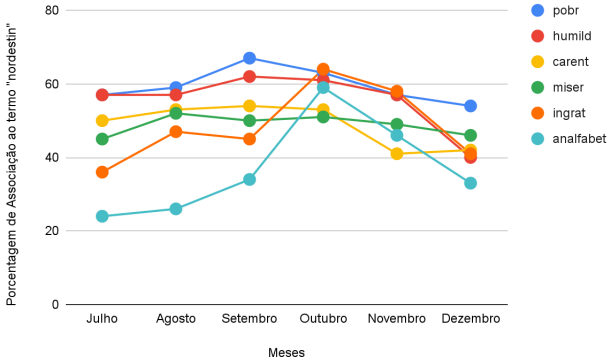


**Figura 11: Gráfico dos principais termos associados a "nordestin" que carregam conotação negativa**

## 6 CONCLUSÃO

Com os resultados provindos dos modelos, há evidências de que à medida que os *tweets* se aproximam dos meses das eleições (Setembro e Outubro), as palavras mais comumente associadas a eles se tornam cada vez mais negativas. Em Setembro, a maioria das palavras negativas que são destaque em Julho, Agosto e Novembro estão no seu ápice de associação e em Outubro, essas palavras

continuam no topo e têm acréscimo de outras palavras de sentido negativo.

Também, é possível notar que nenhuma palavra de sentido positivo está entre as mais associadas ao termo “nordestin”, sendo todas neutras ou negativas. Portanto, é possível concluir que a visão dos usuários do Twitter que comentaram acerca das eleições é, em maior parte, negativa e essa negatividade aumentou à medida que as eleições se aproximavam, associando os nordestinos ao analfabetismo e pobreza, situações que se encaixam dentro do espectro de xenofobia.

Em relação a trabalhos futuros, é planejada utilização de diferentes algoritmos e métodos de análise de texto na pesquisa, como técnicas de análises de sentimento, de classificação textual e o uso de bibliotecas como o Spacy [2], que possui diferentes técnicas de processamento de linguagem natural, categorização textual e modelos de similaridade entre palavras.

## REFERÊNCIAS


[1] GAZETA DO POVO, 2021. Redes sociais como canal de divulgação de conteúdo crescem em meio à pandemia. Disponível em: < https://www.gazetadopovo.com.br/gazz-conecta/redes-sociais-como-canal-de-divulgacao-de-conteudo-crescem-em-meio-a-pandemia/>. Acesso em: 20/01/2023.

[2] O GLOBO, 2022. Mapa da votação. Disponível em: < https://infograficos.oglobo.globo.com/politica/eleicoes-2022/mapa-votacao-municipios-e-estados-do-brasil.html/presidente?desempenho=geral>. Acesso em 20/01/2023.

[3] TECHTUDO, 2019. Conheça as redes sociais mais usadas no Brasil e no mindo em 2018. Disponível em: <https://www.techtudo.com.br/noticias/2019/02/conheca-as-redes-sociais-mais-usadas-no-brasil-e-no-mundo-em-2018.ghtml>. Acesso em 20/01/2023.

[4] MULLAH, Nanlir Sallau; KOTZÉ, Eduan; Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. IEE ACCESS, v.9, p. 88364-88376, 2021

[5] OLUWAFEMI, Oriola; KOTZÉ, Eduan; Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. IEE ACCESS, v. 8, p. 21496-21509, 2020.

[6] Araujo, G. D. de, F. L. de Moraes, e I. T. Pisa. “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: Research Trends, vol. 14, nº 3 - jul-set, agosto de 2020, p. e020006, doi:10.36311/1940-1640.2020.v14n3.10179.

[7] BOYD, Danah M.; ELLISON, Nicole B. 2007. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, Wiley Online Library 13, 1, 210–230.

[8] BARBOSA, Iaan Carvalho. Reconhecimento de mensagens com teor transfóbico no Twitter. SISTEMOTECA - Sistemas de Biblioteca da UFCG, Trabalho de Conclusão de Curso - Artigo - Ciência da Computação. 2021.

[9] LIDDY, Elizabeth D. Natural language processing. 2001.

[10] VIJAYARANI, S., Ms J. Ilamathi, and Ms Nithya. "Preprocessing techniques for text mining-an overview."International Journal of Computer Science Communication Networks v5, n. 1, p. 7-16, 2015.

[11] MARTINEZ, Wendy L.; MARTINEZ, Angel R.; SOLKA, Jeffrey. Exploratory data analysis with MATLAB. Chapman and Hall/CRC, p.3-5, 2017.


[2] Encontrado em https://spacy.io/. Acesso em: 25/01/2023, 22:00