UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Sarah Maria Braga Silva

Análise de sentimentos expressos no *Twitter* em relação aos candidatos da eleição presidencial de 2022

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Sarah Maria Braga Silva

Análise de sentimentos expressos no *Twitter* em relação aos candidatos da eleição presidencial de 2022

Trabalho de conclusão de curso apresentado à Faculdade de Gestão e Negócios da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Gestão da Informação.

Orientador: Elaine Ribeiro de Faria Paiva

Universidade Federal de Uberlândia – UFU

Faculdade de Gestão e Negócios

Bacharelado em Gestão da Informação

Uberlândia, Brasil

2023

Sarah Maria Braga Silva

Análise de sentimentos expressos no *Twitter* em relação aos candidatos da eleição presidencial de 2022

Trabalho de conclusão de curso apresentado à Faculdade de Gestão e Negócios da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Gestão da Informação.

Elaine Ribeiro de Faria Paiva
Orientador

Fabiano Azevedo Dorça

Christiane Regina Soares Brasil

Uberlândia, Brasil
2023

Resumo

Redes sociais são espaços na internet que permitem o compartilhamento e criação de conteúdo pelos seus usuários, sendo consideradas importantes para expressão de opiniões e debates online. Dentre as redes sociais existentes, temos o *Twitter*, cujo foco principal é o compartilhamento de textos curtos, chamados de *tweets*.

Um dos principais assuntos debatidos na rede é a política, principalmente durante períodos de campanha eleitoral, como foi o caso de 2022. Nesse sentido, políticos têm utilizado as redes sociais cada vez mais para se conectar com seus eleitores e divulgar suas campanhas, gerando engajamento, comentários, notícias e discussões a seu respeito.

Este trabalho visa analisar sentimentos expressos pelos usuários do *Twitter* em relação aos candidatos à presidência da eleição de 2022, com o objetivo de verificar se o desempenho dos candidatos na eleição presidencial está relacionado com a sua popularidade nas redes sociais. Para isso, dados provenientes dessa rede social foram coletados, pré-processados e classificados com o *SVM*, o algoritmo classificador escolhido para este projeto.

A partir dos resultados, notou-se que é possível obter semelhanças entre a popularidade do candidato na rede social e sua intenção de voto em pesquisas eleitorais. Porém, não foram encontradas associações entre a popularidade de certo candidato e seu desempenho na eleição. Também foi observado que a quantidade de *tweets* coletados do candidato é uma métrica importante a ser analisada em conjunto com a sua taxa de aprovação, pois apesar de o candidato ser bem aprovado nas redes sociais, nem sempre ele será o mais comentado e, consequentemente, mais conhecido pelos usuários da rede social, impactando diretamente o seu desempenho na eleição.

Palavras-chave: Análise de Sentimentos, Mineração de textos, *Twitter*, Eleições presidenciais.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nos últimos anos, as redes sociais têm se moldado cada vez mais como um espaço para seus usuários debaterem e expressarem suas opiniões sobre assuntos em comum. Aliado ao fato de que essas plataformas geram uma grande quantidade de dados por dia, pesquisadores e empresas conseguem coletá-los para realizar análises de conteúdo em grande escala (ARAÚJO et al., 2013).

Milhares de pessoas utilizam o *Twitter* como canal de expressão pública de suas opiniões sobre diversos assuntos. Por esse motivo, pode ser considerado um bom parâmetro para verificar quais assuntos estão sendo mais comentados naquela região, já que a própria plataforma possui um ranking denominado de *trending topics*, com os termos mais comentados pelos seus usuários naquele momento.

Dentre os diversos assuntos debatidos, temos a política, tema que gera interesse e comentários de várias pessoas. Segundo Rossetto, Carreiro e Almada (2013), o *Twitter* se destaca como um local de divulgação de discussões políticas, possuindo 3 papéis importantes: (1) é uma forma de obter informações políticas rápidas e sem filtro, de diversas fontes diferentes; (2) auxilia no anseio dos usuários que desejam participar ativamente do processo político, como difusores de informação; (3) atua como ferramenta de negócio para cobertura de notícias políticas, assim como políticos que desejam divulgar suas campanhas.

Um exemplo de utilização do *Twitter* como plataforma de discussão política foi no dia 7 de setembro de 2022, marco de 200 anos de independência do país. Nessa data, os *trending topics* do *Twitter* continham apenas termos relacionados a esse evento, como #7desetembro, COM LULA PELO BRASIL, Independência, Viva o Brasil e #BolsonaroMente (FONSECA; SANTINO, 2022).

Além disso, nas manifestações ocorridas em 2013, movidas pelo aumento no preço das passagens de ônibus, as redes sociais se transformaram em um espaço de expressão da democracia, uma vez que as pessoas compartilhavam suas opiniões sobre o ocorrido, difundiam informações e organizavam novas manifestações (CARVALHO, 2020). Desde então, Marques e Sampaio (2011) detectaram um crescimento considerável na utilização de redes sociais no âmbito político no Brasil, principalmente em épocas de campanhas eleitorais.

Ao ser aplicada no ramo da política, a análise de sentimentos possibilita mensurar a popularidade e aprovação de cada candidato, podendo ser utilizada como um possível meio de realizar pesquisas eleitorais (PEREIRA, 2019). Diversos trabalhos têm buscado analisar essa possibilidade, por meio da utilização de algoritmos de classificação. Os *tweets*

coletados são classificados com o sentimento positivo, neutro ou negativo em relação àquele candidato. Assim, aqueles que possuem uma maior quantidade de *tweets* positivos, possuem uma chance maior de se destacaram nas eleições.

Com a chegada das eleições que irão definir o futuro do país pelos próximos anos, aliado ao grande volume de dados gerados nas discussões políticas no *Twitter*, torna-se relevante analisar a opinião dos usuários da rede em relação aos candidatos. Dessa forma, é possível verificar se há uma correlação entre a aprovação dos candidatos nas redes e o resultado das eleições.

Portanto, o problema a ser analisado neste estudo consiste em verificar se o desempenho dos candidatos à presidência do Brasil no ano de 2022 pode ser predito a partir de sua popularidade no *Twitter*.

1.1 Objetivo

1.1.1 Objetivo Geral

Este trabalho tem como objetivo utilizar técnicas de mineração de dados textuais para analisar comentários oriundos da rede social *Twitter*, a fim de verificar se o desempenho de candidatos à presidência da República nas eleições brasileiras de 2022 está relacionado com a sua popularidade nas redes sociais.

Para atingir esse objetivo, serão utilizadas técnicas de pré-processamento e classificação de textos, que irão possibilitar a identificação do sentimento expresso no comentário realizado pelo usuário. Por fim, será feita uma comparação entre os resultados obtidos na classificação de sentimentos e o resultado final das eleições presidenciais.

1.1.2 Objetivos Específicos

- Utilizar uma ferramenta para coleta de comentários do *Twitter* referentes às eleições presidenciais de 2022, entre o período de 01/06/2022 e 01/10/2022, criando assim uma base de dados que poderá ser utilizado por outros trabalhos;
- Classificar o sentimento de cada *tweet* em relação à cada candidato, utilizando algoritmos supervisionados e uma base de dados manualmente rotulada para este fim;
- Investigar o uso do *transfer learning* na tarefa de classificação a fim de que não seja necessário rotular manualmente uma amostra de dados para treinamento do modelo, mas usar uma base rotulada para outro fim;

- Analisar a viabilidade do *Twitter* como ferramenta para predição de eleições, comparando o resultado da classificação de sentimentos com o resultado final das eleições.

1.2 Organização do trabalho

O restante do trabalho está organizado da seguinte maneira.

- **Capítulo 2 - Referencial Teórico:** fundamentação téorica necessária para o desenvolvimento do trabalho, incluindo uma visão geral sobre redes sociais, descrições sobre as etapas necessárias para o desenvolvimento de projetos de mineração textuais e descrições e comentários sobre trabalhos utilizados como base para o desenvolvimento deste projeto;
- **Capítulo 3 - Método para análise entre *tweets* e o resultado das eleições:** discute todas as etapas realizadas para desenvolvimento deste trabalho, com descrição e justificativa das ferramentas utilizadas;
- **Capítulo 4 - Resultados:** discussão dos resultados obtidos com as análises realizadas, e comparação com os dados reais das eleições de 2022;
- **Capítulo 5 - Conclusão:** as principais conclusões do trabalho são apresentadas, assim como sugestões para trabalhos futuros.

2 Referencial Teórico

Este capítulo tem como objetivo apresentar a fundamentação teórica necessária para o desenvolvimento e entendimento deste estudo, além de trabalhos relacionados ao tema.

A primeira seção apresenta uma contextualização das eleições no Brasil, desde as primeiras eleições até os dias atuais. A segunda seção discute a influência das redes sociais nas eleições, apresentando seus principais conceitos e fundamentos que compõe a base deste trabalho. As seções Mineração de dados e Mineração de textos apresentam os principais conceitos que serão aplicados na prática durante o desenvolvimento do projeto. Por último, a seção trabalhos relacionados apresenta alguns trabalhos que serviram como base para a realização deste estudo.

2.1 Eleições no Brasil

Segundo Feloniuk (2014), as primeiras eleições realizadas no Brasil aconteceram nos primeiros anos após a chegada de Portugal. Apesar de nenhum documento ou ata ter sobrevivido ao tempo, sabemos que foi o início de uma tradição que se perpetua até os dias atuais.

Durante o período imperial (1822-1889), a legislação que regulamentava as eleições sofreu muitas modificações, sendo a principal o estabelecimento da Lei Saraiva no ano de 1881 (CAJADO; DORNELLES; PEREIRA, 2014). Além da proibição do voto por analfabetos e criação do título de eleitor, a Lei Saraiva ou Lei do Censo também foi responsável pelo fim das eleições indiretas no país (FARIA, 2013). Dessa forma, os candidatos eram escolhidos não mais por colégios eleitorais, mas sim de forma direta pelos eleitores brasileiros.

Com a instituição da República no ano de 1889, o cenário político não sofreu tamanhas modificações. De acordo com Cajado, Dornelles e Pereira (2014), o período conhecido como Primeira República (1889 - 1930) foi visto como uma idade das trevas eleitoral. A cada eleição, o poder das elites tradicionais continuava perpertuando devido à utilização de artifícios fraudulentos.

Outro marco importante para o cenário político brasileiro foi a redemocratização do país, após o fim da Ditadura Militar. Com a promulgação da Constituição de 1988 e implementação de direitos como voto secreto e universal, além da legalização de partidos socialistas, o país implementou novamente a base para sua democracia, que vigora até os dias atuais (MELO, 2010).

O cenário político atual é marcado por uma polaridade de partidos e seus candidatos, em específico PT e PSDB, que dominam não só as eleições presidenciais, como também as regionais (BORGES, 2015). O estudo de Reis (2014) aponta que as manifestações ocorridas em 2013, marcadas pela forte tensão entre os dois partidos principais (PT e PSDB), além de conflitos ideológicos entre direita e esquerda, tiveram consequências imediatas nas eleições seguintes. Em 2014, Dilma Rousseff (PT) foi eleita com 51,64% dos votos válidos, contra 48,36% de seu concorrente Aécio Neves (PSDB). Já em 2018, Jair Bolsonaro (PSL) foi eleito com 55,13% dos votos válidos, contra 44,87% de seu concorrente Fernando Haddad (PT) (TSE, 2022).

Outro elemento importante no cenário político atual são as pesquisas pré-eleitorais. Apesar de serem realizadas desde 1942, quando o primeiro instituto de pesquisas no Brasil foi criado, elas ganharam mais destaque atualmente por serem consideradas uma ferramenta indispensável no planejamento das campanhas eleitorais (MIGUEL; TOKARSKI; MOTA, 2011). Em seu trabalho, Rossini et al. (2016) cita que a pesquisa eleitoral também é uma importante fonte de informações para eleitores, sendo capaz inclusive de influenciar suas decisões. Os eleitores indecisos, por exemplo, tendem a votar em candidatos que estão à frente das pesquisas, como uma forma de não desperdiçar seu voto.

De acordo com Mauro Paulino, ex-diretor geral do Datafolha, um dos principais órgãos responsáveis por realizar pesquisas eleitorais no país, o primeiro passo para realização da pesquisa é a elaboração do questionário, com as perguntas que devem ser realizadas para os entrevistados. Logo após, é realizado a amostragem, ou seja, escolha do grupo de pessoas que possuem características suficientes para representar toda a população do país. Além disso, para cada grupo, é escolhido pessoas aleatórias para compô-lo. Dessa forma, se a população brasileira é composta de 51% de mulheres, a pesquisa também deverá ser composta de 51% de mulheres escolhidas aleatoriamente.

Ainda segundo Mauro Paulino, existem duas formas de realização da entrevista. A primeira é feita presencialmente, na qual cada entrevistado recebe um cartão circular, para que a ordem dos nomes não interfira na escolha do entrevistado. A outra forma de realização da pesquisa, é por telefone. Nesse caso, a ordem de fala de cada candidato é feita de forma aleatória, de modo a minimizar a influência na escolha do entrevistado.

Após a realização das entrevistas, os dados são analisados a fim de chegar na porcentagem de intenção de voto em cada candidato e a margem de erro de cada um, para depois ser registrada na Justiça eleitoral e por fim publicada (SOUZA, 2022).

2.2 Eleições nas redes sociais

Desde o seu surgimento, a Internet tem possibilitado uma constante exposição de pontos de vista e troca de opiniões entre pessoas. Seus primeiros meios de comunicação,

como fóruns e blogs pessoais, já possibilitavam essa troca de opiniões de maneira personalizada, porém, foi com a criação das redes sociais que esse movimento se intensificou (TEIXEIRA; AZEVEDO, 2011).

Nos últimos anos, as redes sociais têm sido também utilizadas para debates e discussões políticas. Além disso, vários trabalhos tem as usado para entender melhor a aceitação de um dado candidato por parte da população.

2.2.1 Redes sociais

De acordo com Zenha (2018), redes sociais são um ambiente digital feito através de uma interface visual própria, com o objetivo de conectar perfis que possuam afinidades, pensamentos e interesses em um tema em comum. Mais de dois terços da população online global é ativa em alguma rede (INSPER, 2022). Essa popularidade está associada com o principal objetivo de uma rede social, que é permitir que seus usuários criem e compartilhem conteúdos de seu interesse.

Em seu trabalho, Benevenuto, Almeida e Silva (2012), definem as principais funcionalidades presentes nas redes sociais, sendo elas:

- Perfis de usuários: identificam o indivíduo no sistema, mostrando uma descrição de informações consideradas relevantes para o usuário como interesses e fotos;
- Atualizações: formas efetivas de ajudar usuários a descobrir conteúdos;
- Comentários: usuários podem comentar conteúdos de outros usuários; e
- Avaliações: usuários podem avaliar conteúdos divulgados por outros usuários.

Dentre as redes sociais existentes atualmente, se destaca o *Twitter*[1], por possuir 1,3 bilhão de usuários, os quais enviam mais de 500 milhões de *tweets* ao todo por dia (AHLGREN, 2022). Desenvolvido no ano de 2006, no formato de micro-blogging, o *Twitter* é uma rede social onde as pessoas podem publicar textos com conteúdos variados de até 280 caracteres, conhecidos como *tweets* (RECUERO; ZAGO, 2016).

A facilidade de transmitir informações e opiniões é um dos principais motivos pelos quais o *Twitter* se destaca frente às outras redes existentes (NASCIMENTO; OSIEK; XEXÉO, 2015). Seus usuários não utilizam apenas para divulgar opiniões pessoais, mas também como uma forma de acompanhar informações sobre fatos e eventos em geral. Ainda segundo o autor, é justamente por isso que o *Twitter* se torna uma importante fonte de opiniões sobre eventos que podem ser analisados posteriormente, por meio de técnicas como Análise de Sentimentos, e utilizados em várias áreas como política e *marketing*.

[1] https://twitter.com/

Um dos exemplos de uso das redes sociais em questões políticas foi o caso do marco de 200 anos de independência do Brasil. Nesse dia, os usuários do Twitter utilizaram termos como #7desetembro, COM LULA PELO BRASIL, Independência, Viva o Brasil e #BolsonaroMente para debater sobre a situação política atual do país(FONSECA; SANTINO, 2022).

2.2.2 Marketing político

Segundo Alves (2018), marketing político pode ser definido como um conjunto de técnicas e procedimentos realizados a fim de aumentar a visibilidade do candidato, mostrando-o diferente e melhor do que os demais.

Devido à nova realidade proporcionada pelas redes sociais e a tecnologia, no ano de 2018 o TSE (Tribunal Superior Eleitoral) autorizou a propaganda eleitoral na Internet. Desde então, candidatos e partidos políticos utilizam as redes sociais como um meio de divulgação de suas campanhas e atos políticos realizados, como eventos ou até mesmo manifestações (SILVA; SANTOS, 2020).

O crescente uso das redes sociais cria uma necessidade nos atores políticos de estarem atentos aos anseios e necessidades da população (LEMOS, 2019). Esse fenômeno de observação de necessidades para influenciar o comportamento foi aplicado por Donald Trump em 2016. Seu sucesso nas eleições foi atribuído às campanhas realizadas nas redes sociais pela empresa *Cambridge Analytica*, especializada em influenciar o comportamento de pessoas com base em coleta de dados nas redes sociais (FLORES, 2018).

Por outro lado, é cada vez mais frequente discussões sobre a presença de *fake news* nas redes sociais. Tendo como definição a produção e distribuição de notícias falsas, as *fake news* possuem o poder de influenciar eleitores, seja favorecendo ou desfavorecendo candidatos específicos (ALMEIDA, 2018). No Brasil, esse tema ganhou destaque nas eleições de 2018, por causa da grande quantidade de informações falsas identificadas na mídia e redes sociais. Segundo Dourado (2020), o até então candidato à presidência Jair Bolsonaro, e vencedor das eleições de 2018, foi o maior beneficiado pela distribuição de *fake news*, quer seja direta ou indiretamente.

Os robôs sociais, ou *bots*, são os maiores aliados e responsáveis pela distribuição de notícias falsas. De acordo com Jardelino, Cavalcanti e Toniolo (2021), eles são utilizados já há algum tempo, principalmente na rede social *Twitter*. Um estudo realizado pela Fundação Getúlio Vargas em 2017 mostra que essa distribuição atinge diretamente os processos políticos, por causa das influências que possuem na opinião pública, quer seja criando falsos cenários ou manipulando os assuntos do momento no *Twitter*.

Além disso, segundo Filho, Almeida e Pappa (2014), as redes sociais, em específico o *Twitter*, também podem ser utilizadas para realização de pesquisas eleitorais e predição

de resultados de campanhas. Em seu trabalho, o autor coletou mensagens referentes aos candidatos de cidades selecionadas previamente, removendo *spammers* e conteúdos jornalísticos, e depois contabilizou os votos conforme o sentimento identificado nas mensagens analisadas. Como resultado, chegou à conclusão de que análises como essa são capazes de melhorar os números alcançados nas tradicionais pesquisas eleitorais.

Devido ao alto volume de dados, empresas e pesquisadores que analisam dados oriundos das redes sociais, necessitam de técnicas específicas para processá-los a fim de retirar informações valiosas. A *Cambridge Analytica* por exemplo, citada anteriormente, utiliza técnicas e softwares de *big data* e mineração de dados a fim de analisar o perfil do público de cada candidato, para que dessa forma, estratégias mais efetivas e personalizadas possam ser desenvolvidas.

2.3 Mineração de dados

Mineração de dados é o processo de extrair ou minerar padrões, modelos ou outros tipos de conhecimento de grandes volumes de dados (HAN; PEI; TONG, 2022). Sua aplicação envolve diversas etapas, como por exemplo, a preparação dos dados afim de obter padrões, ou seja, subconjuntos que descrevam algum tipo de comportamento observado. Além disso, é necessário que esses padrões tenham algum grau de certeza e forneçam informações até então desconhecidas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

De acordo com López (2021), o projeto CRISP-DM foi desenvolvido no ano de 1996, sendo um modelo do processo de mineração de dados, o qual diz que o ciclo de vida de um projeto na área consiste de 6 etapas, como exemplificado na Figura 1. Sua sequência não é obrigatória, podendo ocorrer voltas à fases anteriores para revisão do trabalho já realizado.

A primeira etapa consiste na compreensão do negócio da empresa e os principais objetivos a serem alcançados de acordo com essa perspectiva. Com os objetivos traçados, os profissionais responsáveis pelo processo deverão analisar a situação, discutindo sobre os riscos que irão enfrentar, além de determinar os custos e tecnologias aplicadas no projeto. Essa fase é considerada crucial para o processo de mineração de dados, pois é necessário garantir que o modelo desenvolvido atende aos verdadeiros problemas que a empresa possui.

A segunda etapa do processo consiste na definição de quais e como serão coletados os dados a serem utilizados no projeto. É necessário ganhar familiaridade com os dados, entender os problemas de qualidade, como atributos faltantes ou dados nulos, e explorá-los a fim de descobrir *insights* iniciais.

A terceira etapa consiste nas atividades relacionadas à construção da base de dados

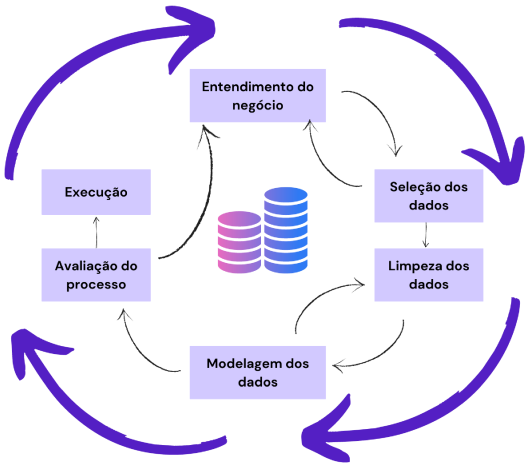

Figura 1 – Etapas do CRISP-DM. Adaptado de Chapman et al. (2000)

que será aplicada no modelo. O tempo gasto nessa fase compreende 50 a 70% do projeto, e consiste em etapas como: seleção dos dados, limpeza dos dados através de algoritmos de pré-processamento, construção dos dados, caso seja necessário acrescentar atributos, e integração dos dados, caso exista mais de uma fonte.

A quarta etapa consiste na modelagem do algoritmo. Neste momento são testados vários modelos e técnicas diferentes afim de analisar qual o melhor para o problema em questão. Além disso, é nesse momento também que pode ser necessário voltar para a etapa de preparação dos dados, pois diferentes modelos exigem formatos de dados diferentes.

Logo após a construção do modelo, tem-se a avaliação dos resultados obtidos. Essa etapa geralmente exige modificações nas etapas anteriores, caso o resultado não tenha atingido as expectativas iniciais. Neste momento, é importante avaliar se todos os objetivos e problemas traçados na primeira etapa foram atingidos, e quais serão as próximas etapas do projeto.

Por último, o *deploy* é realizado, momento no qual o modelo é colocado em produção e apresentado para todos os usuários. A complexidade da etapa depende do acordo firmado com a empresa, podendo ser simples como a geração de um relatório, ou complexo como a implementação de um processo automático de mineração de dados para toda a empresa.

2.4 Mineração de textos

Os conceitos e técnicas da mineração de texto foram originados da mineração de dados, podendo ser inclusive visto como uma extensão da área (MORAIS; AMBRÓSIO,

2007). Segundo Passos (2006), a mineração de texto é um conjunto de métodos usados para navegar, organizar e analisar informações em bases textuais. Ao contrário dos mecanismos de buscas, que simplesmente processam as informações recebidas, as técnicas utilizadas na mineração têm como objetivo descobrir informações até então desconhecidas.

A Internet possui mecanismos, como redes sociais e fóruns, que facilitam a expressão do ponto de vista e opinião do usuário. Nesse sentido, a mineração de textos é uma aliada estratégica para aqueles que buscam compreender a opinião de usuários sobre um determinado assunto (FIGUEIREDO; CATINI; MENDES, 2018). O ramo da mineração que tem como objetivo analisar opiniões, sentimentos, emoções e atitutes expressadas em textos é chamado de Análise de Sentimentos (MEDHAT; HASSAN; KORASHY, 2014) .

Segundo o trabalho de Passos (2006), os principais desafios encontrados na mineração de textos giram em torno da alta dimensionalidade dos dados, ambiguidade e dados ruidosos encontrados. Quando aplicado em redes sociais, esses problemas se intensificam por causa da grande presença de gírias e sarcasmo nos textos.

Assim como a Mineração de dados, a Mineração de textos pode ser dividida em várias etapas. Em seu trabalho, Aranha (2007) definiu as 6 etapas principais, exemplificadas na Figura 2.

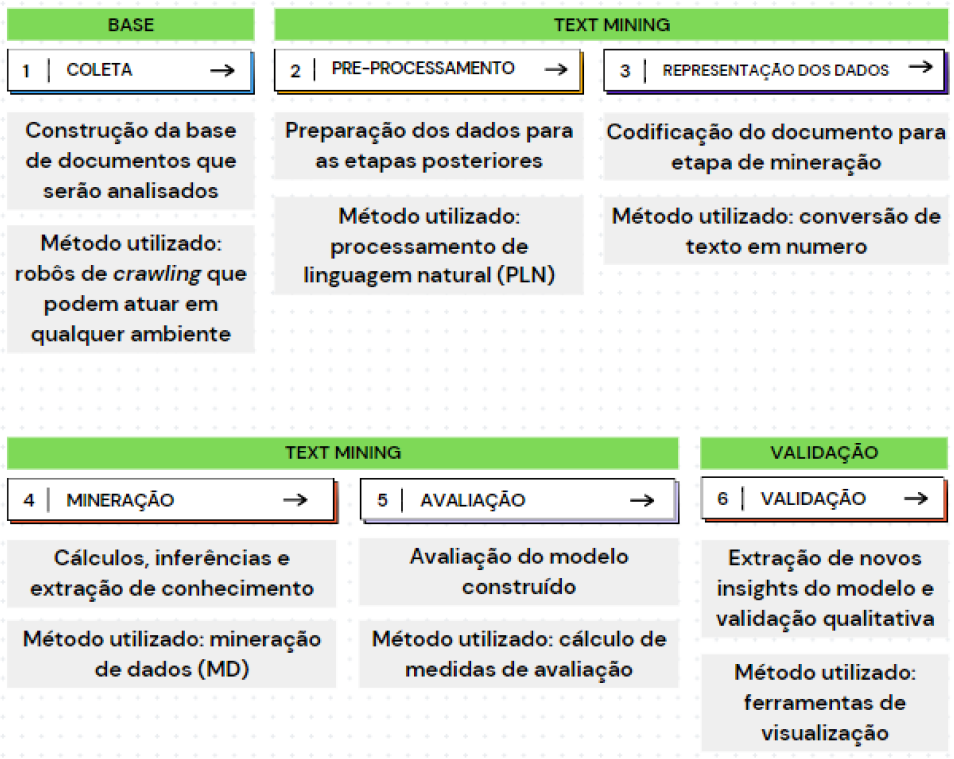

Figura 2 – Etapas da Mineração de Texto. Adaptado de Aranha (2007)

2.4.1 Coleta de dados

A etapa de coleta de dados é o processo de busca e recuperação dos dados a serem utilizados na análise, ou seja, aqueles dos quais se pretende extrair algum tipo de informação (CHAPMAN et al., 2000).

De acordo com Benevenuto, Almeida e Silva (2012), um dos métodos mais comuns de extração de dados textuais é através de APIs de redes sociais. No contexto *web*, uma API é um conjunto de requisições HTTP, utilizadas para extrair informações de alguma plataforma externa (SALVADORI et al., 2015). Ao realizar a requisição por meio de APIs, os dados são retornados em sua maioria no formato XML e JSON, facilitando o seu uso em diversas outras aplicações e linguagens de programação.

Ainda segundo Benevenuto, Almeida e Silva (2012), várias plataformas oficiais como *Twitter*, *Flickr*, *Youtube*, dentre outras, possuem APIs que facilitam a coleta e análise de dados de seus usuários.

2.4.2 Pré-processamento

O pré-processamento consiste na aplicação de técnicas de transformação de dados, a fim de obter um formato adequado para construção de algoritmos (MARTINS; MONARD; MATSUBARA, 2003).

Segundo Junior (2007), a técnica conhecida como *tokenização* tem como finalidade extrair unidades mínimas de texto a partir de um texto livre. Na grande maioria das vezes, essa unidade mínima, denominada de *token*, está relacionada a cada palavra de uma frase. Ainda segundo o autor, a obtenção de cada *token* é feita por meio da quebra do texto em diferentes delimitadores, sendo o padrão o "espaço", porém, outros comuns são: ()<>!?.;'-. Apesar de parecer uma técnica simples, os delimitadores podem exercer diferentes papéis dentro da mesma frase. Um exemplo é o "ponto", que pode ser usado tanto para marcar o fim de uma frase, quanto para abreviações.

Stopwords são termos que podem ser considerados irrelevantes para a tarefa de mineração de textos, como preposições, artigos e conjunções. Segundo Martins (2003), quando eliminamos esses termos, conseguimos obter uma drástica redução no conjunto de atributos, e a análise não é influenciada por palavras irrelevantes que aparecem em alta frequência no texto.

Outra técnica de pré-processamento bastante utilizada é o algoritmo de *stemming*, o qual consiste em uma normalização linguística, na qual as formas variantes de um termo são reduzidas a uma forma comum (MARTINS, 2003). Assim, palavras como representar, representado e representando, aparecem após a aplicação do *stemming* como represent e são interpretados como tendo o mesmo significado. Apesar de ser bastante utilizado, o algoritmo deve ser aplicado com cautela, pois como demonstra Willett (2006) em seu

estudo, palavras com significados diferentes podem ser reduzidas ao mesmo termo, e no final serem vistas como iguais.

2.4.3 Representação de dados

O processo de representação dos dados tem como principal objetivo codificar o documento de forma que facilite sua posterior manipulação e análise. De acordo com Junior (2007), as duas principais representações para documentos são o TF-IDF e a codificação *bag of words*.

O TF-IDF é a combinação das medidas *term frequency* (*tf*) e *inverse document frequency* (*idf*). Segundo Martins, Monard e Matsubara (2003), a medida *term frequency* utiliza o número de ocorrências do termo t no documento d. No entanto, quando termos com alta frequência aparecem na maioria dos documentos, eles nao fornecem informações úteis para diferenciar documentos. Por outro lado, a medida *inverse document frequency* diminui a importância dos termos com alta frequência, variando inversamente ao número de documentos x, que contém o termo t, em uma coleção de documentos. Dessa forma, a medida TF-IDF é a combinação das medidas *tf* e *idf*.

A codificação *bag of words*, segundo Junior (2007), é aquela na qual cada documento é visto como diversos *tokens* agrupados. Cada *token* representa um termo diferente do documento e seu valor é obtido por meio da frequência de ocorrências no documento.

2.4.4 Mineração

A aplicação dos algoritmos de mineração é fundamental para o processo de Mineração de Textos, pois são eles que possibilitam a descoberta de informações até então desconhecidas. Para o objetivo proposto neste trabalho, faz-se necessário a aplicação de algoritmos de classificação, a fim de rotular os dados colhidos em 3 classes distintas: positivo, negativo e neutro.

De acordo com Han, Pei e Tong (2022), classificação é o processo de encontrar um conjunto de funções que descrevem e diferenciam classes, com o objetivo de utilizar o modelo para predizer a classe de objetos não classificados. Para isso, o modelo baseia-se na análise prévia de um conjunto de dados pré-rotulados. Dessa forma, o conjunto de dados de entrada é composto dos atributos da instância e seu respectivo rótulo.

Dentre os algoritmos de classificação utilizados na Mineração de textos, se encontra o *Naive Bayes*. Segundo Rish et al. (2001), o *Naive Bayes* é um classificador probabilístico, baseado no Teorema de Bayes, portanto, a classificação é realizada por meio do cálculo da probabilidade da instância pertencer a uma classe ou outra, com base no conhecimento prévio das condições do evento. Além disso, é importante destacar também que, por

ser baseado no Teorema de *Bayes*, o modelo supõe que exista uma independência nas características das instâncias.

Por outro lado, segundo o trabalho de Alves et al. (2014), outro algoritmo que se destaca na classificação de textos por seu bom desempenho é o *SVM*. Segundo o autor, o algoritmo busca traçar linhas de separação entre classes distintas, analisando os dois pontos mais próximos de outras classes, a fim de encontrar aquela que se distancia mais de cada classe. Após descoberta essa reta, o programa conseguirá predizer a qual classe pertence a nova instância ao checar de qual lado da reta ela está.

2.4.5 Avaliação do modelo

Uma das definições a serem feitas durante essa etapa é a escolha do método de avaliação. Dentre os mais utilizados se destaca o *cross-validation*, cujo principal objetivo é mensurar o poder de generalização do algoritmo, assim como prevenir problemas como *overfitting*[2] (BERRAR, 2018).

Neste estudo, foi utilizado o método *k-fold cross-validation*. Segundo Fushiki (2011), esse método consiste na divisão da base de dados em K partes. A cada execução, 1 parte é utilizada como teste e as demais para treino do modelo, sendo que ao final de cada execução, são calculadas as medidas de avaliação desejadas. Dessa forma, o valor final de cada medida de avaliação é a média dos valores obtidos em cada execução.

Em problemas de classificação binária, é comum denominarmos a classe alvo do estudo como positiva, geralmente minoritária, sendo a classe restante portanto denominada de negativa (ANACLETO, 2010). Outra definição a ser feita na análise de indicadores do modelo, é a escolha das medidas de avaliação a serem utilizadas. Para isso, utiliza-se o cenário de cada instância após a classificação, sendo possível as seguintes opções, considerando o cenário em que o classicador é treinado utilizando as classes "Positivo", "Negativo":

- Verdadeiro positivo (VP): são exemplos da classe positiva (alvo) que foram previstos corretamente;
- Falso positivo (FP): são exemplos da classe positiva (alvo) que foram previstos incorretamente;
- Verdadeiro negativo (VN): são exemplos da classe negativa (não-alvo) que foram previstos corretamente;
- Falso negativo (FN): são exemplos da classe negativa (não-alvo) que foram previstos incorretamente.

[2] Ocorre quando o modelo não tem capacidade de generalização, ou seja, se ajusta muito bem aos dados de treino, porém se mostra ineficaz em outros dados

Segundo Schneider (2018), acurácia pode ser definida como a proporção dos verdadeiros positivos , com relação ao total de previsões. A Equação 2.1 apresenta a fórmula utilizada para o cálculo da métrica.

$$\mathrm{A} = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

A precisão pode ser definida como a proporção dos verdadeiros positivos obtidos em relação ao total de verdadeiros positivos e falsos positivos (JUNIOR et al., 2022). A Equação 2.2 apresenta o cálculo da métrica.

$$\mathrm{P} = \frac{VP}{VP + FP} \tag{2.2}$$

Outra métrica bastante utilizada é o *revocação*, que pode ser definida como a proporção dos verdadeiros positivos em relação ao total de verdadeiros positivos e falsos negativos (SCHNEIDER, 2018). A Equação 2.3 apresenta o cálculo para o *recall*.

$$\mathrm{R} = \frac{VP}{VP + FN} \tag{2.3}$$

É importante destacar contudo, que este estudo se trata de um problema de classificação multiclasse, pois temos mais de 2 classes, sendo elas Positiva, Negativa e Neutra. Dessa forma, é preciso utilizar o cálculo macro das medidas apresentadas anteriormente. Assim, cada indicador será calculado 3 vezes, considerando em cada uma delas uma classe diferente como Positiva e as demais como Negativa. No final, o valor do indicador-macro é a média dos indicadores calculados anteriormente (GRANDINI; BAGLI; VISANI, 2020).

2.4.6 Validação do modelo

Essa etapa diz respeito à validação qualitativa do modelo construído, ou seja, verificar se os objetivos traçados no início do projeto foram atendidos, e se o modelo é realmente capaz de extrair informações novas e confiáveis a partir dos dados disponíveis (JUNIOR, 2007).

Também conhecida por pós-processamento, nessa etapa são construídos elementos gráficos por meio de ferramentas de visualização, para auxiliar na compreensão dos resultados gerados pelo modelo (GOLDSCHMIDT; PASSOS, 2005). Ainda segundo o autor, técnicas de visualização de dados estimulam a percepção e a inteligência humana, podendo proporcionar a associação de novos padrões por exemplo.

2.5 Trabalhos relacionados

Esta seção tem como objetivo apresentar alguns dos principais trabalhos relacionados com o tema de análise de sentimentos em processos eleitorais. Estes trabalhos foram utilizados como referência para o desenvolvimento deste projeto, e sua busca foi realizada no Google Acadêmico por meio do texto "analise sentimento eleições". A busca inicial retornou 24.500 trabalhos relacionados com os termos buscados, porém foram escolhidos em uma primeira etapa somente aqueles que possuem data de publicação superior ao ano de 2014 e que estejam em português, sendo portanto 15.600 trabalhos. Dentre esses, foram considerados apenas os mais relevantes, ou seja, os 20 primeiros trabalhos retornados na busca anterior. Após a leitura, foram excluídos aqueles que não apresentavam descrições e informações suficientes sobre o algoritmo ou métodos utilizados. Além disso, dentre os restantes, foi selecionado um trabalho por algoritmo, a fim de analisar uma maior gama de métodos dentre os trabalhos presentes. Os trabalhos restantes foram escolhidos e descritos a seguir.

Cristiani, Lieira e Camargo (2020) apresentam um estudo referente às eleições de 2018, uma das áreas de aplicação da análise de sentimentos. Seu principal objetivo foi analisar uma possível relação entre opiniões expressas em redes sociais, no caso o *Twitter*, e o resultado das eleições. Em relação à base de dados analisada, os autores utilizaram a API oficial da rede social para coleta de *tweets* em 8 eventos distintos durante a campanha eleitoral, resultando em uma base com 903.518 *tweets*. Após a coleta, os dados foram pré-processados usando técnicas como padronização, tokenização, remoção de *stopwords*, lematização e TF-IDF. No desenvolvimento do modelo, os autores inicialmente testaram 2 algoritmos diferentes, SVM e *Naive Bayes*. Após a implementação e análise de desempenho, o SVM se destacou em todas as medidas utilizadas, tendo o melhor desempenho obtido por meio da revocação, com 71,05%. Após a análise dos resultados, os autores chegaram a conclusão de que o *Twitter* é uma ótima fonte de pesquisa sobre opiniões, pois o candidato que obteve o maior número de votos, 55%, também foi aquele que obteve uma maior quantidade de *tweets* com sentimento positivo, 37,23%.

Em seu trabalho, Dutra e Francisco (2018) buscaram compreender a influência do marketing durante o período eleitoral. No trabalho em questão, os autores coletaram 1.204.036 *tweets* utilizando os nomes dos candidatos, *hashtags* e palavras como *eleicoes* e *eleicoes2018*. Após a coleta, foram aplicados métodos de pré-processamento como tokenização, remoção de *stopwords* e normalização morfológica. O algoritmo escolhido pelos autores para classificação foi o *Naive Bayes*, o qual obteve 85% de acurácia, 90% de precisão e 84% de revocação. Em sua conclusão, os autores apontaram uma maior classificação de *tweets* como positivos para Bolsonaro, com 34% do total de mensagens positivas, sendo ele também o vencedor das eleições em 2018.

Attux (2017) apresenta um estudo de técnicas referentes à análise de sentimentos

em redes sociais, com o objetivo de realizar a predição dos resultados das eleições para presidência de 2014. Para a construção da base de dados, o autor utilizou como fonte a rede social *Twitter*, coletando dados diariamente durante a campanha eleitoral. No total, foram coletados 330.881 tweets, contendo as principais informações a serem utilizadas na análise como, o texto publicado, localidade, candidato a qual o usuário se refere e data e hora da publicação. Um aspecto interessante do trabalho foi a utilização da ferramenta *Sentiment140*, a qual permite descobrir, por meio de um algoritmo de classificação próprio da plataforma, o sentimento presente no texto publicado. Como resultado da pesquisa, foi possível concluir que o algoritmo de predição se aproximou bastante do resultado da eleição, com exceção de alguns candidatos específicos como Dilma Rousseff. Essa discrepância segundo o autor tem como causa o desbalanceamento dos *tweets* coletados em cada região do país. Assim, a previsão realizada para aqueles candidatos que possuem predominância em determinadas regiões pouco representativas numericamente na base, se torna tendenciosa.

No artigo, Queiroz e Almeida (2020) propuseram uma metodologia de análise de sentimentos que realiza a extração, tratamento e classificação dos *tweets* coletados, identificando seu grau de polaridade a partir da aplicação de dicionários para classificação das palavras em positivas ou negativas. Os dados coletados referem-se ao período entre 01 de agosto e 6 de outubro de 2018, totalizando 88 atributos (características associadas aos *tweets* como usuário e data da publicação) e 4.608 instâncias. Para o processo de mineração, a classificação dos textos como positivos, negativos ou neutros foi realizada por meio da utilização de pacotes de dicionários léxicos. Após a classificação, os autores conseguiram perceber relações diretas entre os comentários da rede social e pesquisas de intenção de voto realizadas pelas principais empresas do país. Posteriormente, foi aplicado o algoritmo K-means para identificação de grupos de palavras mais utilizadas. Dessa forma, os autores conseguiram perceber também os principais temas de discussão dos candidatos.

Na Tabela 1, é apresentada uma comparação entre os trabalhos analisados e suas principais características.

2.6 Considerações finais

A partir dos estudos realizados e apresentados neste capítulo, observa-se a importância que as redes sociais possuem atualmente, em especial o *Twitter*. Nota-se que quando utilizadas em conjunto com algoritmos de mineração de textos, os dados coletados de cada rede fornecem *insights* únicos de seus usuários e suas opiniões, podendo ser analisados com diversos propósitos, sendo um deles a política, tema abordado neste projeto.

Trabalho	Pré processamento	Algoritmo	Medidas de avaliação	Base de dados
(CRISTIANI; LIEIRA; CAMARGO, 2020)	Padronização, Tokenização, Remoção de stopwords, Lematização	Naive Bayes SVM	Acurácia Precisão Recall F-measure	903.518 tweets coletados em 8 eventos/debates distintos durante a campanha eleitoral
(ATTUX, 2017)	Remoção de links Remoção de caracteres especiais Tokenização Tradução	Sentiment140	-	330.881 tweets coletados diariamente entre os dias 27 de agosto e 03 de outubro de 2014
(DUTRA; FRANCISCO, 2018)	Tokenização, Remoção de stopwords, Normalização morfológica	Naive Bayes	Acurácia Precisão Recall	1.204.036 tweets coletados entre 24 a 30 de Junho de 2018
(QUEIROZ; ALMEIDA, 2020)	Padronização, Tokenização, Remoção de stopwords, Lematização	LexiconPT K-means	Método Elbow	4.608 tweets coletados entre 01 de agosto a 6 de outubro de 2018

Tabela 1 – Trabalhos relacionados

Pontua-se também que diversos trabalhos, como aqueles apresentados na seção anterior, têm analisado a viabilidade de realização de pesquisas eleitorais por meio de análises das opiniões dos usuários de redes sociais, em especial o *Twitter*. Assim como observado anteriormente, estes trabalhos concluem que há uma relação direta entre os comentários realizados nas redes sociais e a intenção de voto dos eleitores, aumentando cada vez mais a importância da realização de pesquisas como esta.

3 Método para análise entre *tweets* e o resultado das eleições

Este capítulo tem como objetivo apresentar os procedimentos metodológicos utilizados neste estudo, que visa analisar se o desempenho dos candidatos nas eleições presidenciais está relacionado com sua popularidade no *Twitter*.

A Figura 3 apresenta as etapas do método utilizados neste estudo, sendo que as próximas seções irão explicar mais detalhadamente cada etapa presente na figura.

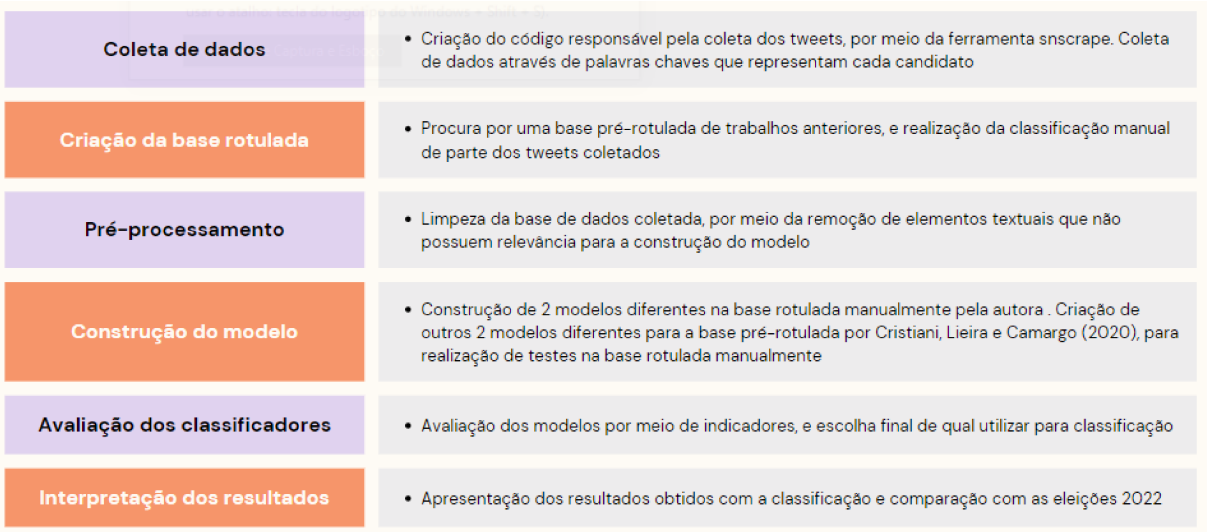

Figura 3 – Metodologia empregada

3.1 Coleta de dados

Para a coleta de dados da plataforma *Twitter* utilizou-se a ferramenta *snscrape*, disponível como uma biblioteca *Python*, que permite extrair *tweets* de um período específico a partir de palavras chaves. Embora o próprio *Twitter* disponibilize uma API[1] oficial para coleta de dados, a mesma possibilita uma extração de no máximo 500 *tweets* por requisição, dificultando dessa forma a coleta de uma quantidade maior de dados.

Por outro lado, a biblioteca *snscrape*, lançada em 8 de julho de 2020, não possui limitações em relação ao número de *tweets* que podem ser coletados em uma mesma requisição, além de ser facilmente utilizada. Outro benefício da ferramenta é a abrangência que possui em relação à extração de dados em redes sociais, podendo ser utilizada não somente no *Twitter*, mas também em redes como *Facebook*, *Instagram* e *Reddit*, por exemplo.

[1] https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

Para a instalação foi necessário utilizar o pip[2] através do comando "pip install snscrape". Depois, a importação da biblioteca por meio do comando "import snscrape.modules.twitter"foi realizada. Diversos argumentos podem ser passados como parâmetros na busca para refiná-la. Aqueles que foram utilizados neste trabalho são descritos a seguir. É importante destacar que a palavra chave a ser buscada não possui um parâmetro específico, mas deve ser passada no início do comando, como mostrado na Figura 4.

- ***since***: indica a data inicial a ser coletada;
- ***until***: indica a data final a ser coletada;
- ***near***: indica o local no qual os *tweets* da busca devem estar próximos. Neste estudo, os locais filtrados foram as capitais dos estados brasileiros, bem como a capital do país;
- ***lang***: indica o idioma dos *tweets*, no caso deste estudo, português.

```python
import snscrape.modules.twitter as sntwitter
import pandas as pd

dat_ini = '2022-8-01'
dat_fim = '2022-8-31'

def search_bolsonaro_sudeste(text = 'bolsonaro', start = dat_ini, end = dat_fim):
    tweets =[]
    for i, tweet in enumerate(sntwitter.TwitterSearchScraper(f'{text} since:{start} until:{end} near:"Belo Horizonte" lang:pt').get_items()
        tweets.append([tweet.date, tweet.id, tweet.content, tweet.username])
    tweets_bolsonaro_bh = pd.DataFrame(tweets, columns = ['date', 'tweet id', 'content', 'username'])

    for i, tweet in enumerate(sntwitter.TwitterSearchScraper(f'{text} since:{start} until:{end} near:"São Paulo" lang:pt').get_items()):
        tweets.append([tweet.date, tweet.id, tweet.content, tweet.username])
    tweets_bolsonaro_sp = pd.DataFrame(tweets, columns = ['date', 'tweet id', 'content', 'username'])

    for i, tweet in enumerate(sntwitter.TwitterSearchScraper(f'{text} since:{start} until:{end} near:"Rio de Janeiro" lang:pt').get_items()
        tweets.append([tweet.date, tweet.id, tweet.content, tweet.username])
    tweets_bolsonaro_rj = pd.DataFrame(tweets, columns = ['date', 'tweet id', 'content', 'username'])

    for i, tweet in enumerate(sntwitter.TwitterSearchScraper(f'{text} since:{start} until:{end} near:"Vitória" lang:pt').get_items()):
        tweets.append([tweet.date, tweet.id, tweet.content, tweet.username])
    tweets_bolsonaro_vt = pd.DataFrame(tweets, columns = ['date', 'tweet id', 'content', 'username'])

    # Junção dos tweets das cidades em 1 arquivo só
    df_final_bolsonaro_sudeste = pd.concat([tweets_bolsonaro_bh, tweets_bolsonaro_sp, tweets_bolsonaro_rj, \
    tweets_bolsonaro_vt], ignore_index=True)
    df_final_bolsonaro_sudeste.to_csv('df_bolsonaro_sudeste_outubro.xlsx')
```

Figura 4 – Exemplo do código utilizado para coleta de dados referente ao candidato Jair Bolsonaro no mês de agosto.

É importante destacar que o parâmetro *near* possui algumas ressalvas. No ano de 2019, o recurso de geolocalização do *Twitter* foi desativado, pois segundo a empresa, a maioria dos usuários não marcava a localização exata nos *tweets* (SOARES, 2019). Dessa

[2] *pip* é um instalador e gerenciador de pacotes Python

forma, não é possível capturar o local exato das publicações na plataforma. Apesar disso, a biblioteca snscrape, de acordo com sua documentação[3], captura os *tweets* que possuem localização marcada no perfil ou no próprio *tweet*, e aqueles que não possuem localização são descartados.

Durante a etapa de coleta de dados, foram realizadas cinco consultas por candidato, sendo cada uma delas referente a cada região analisada. As palavras-chave utilizadas estão descritas na Tabela 2.

Tabela 2 – Palavras-chave utilizadas na busca de *tweets* de cada candidato

Candidato	Palavras-chave
Jair Bolsonaro (PL)	bolsonaro
Lula (PT)	lula
Ciro Gomes (PDT)	ciro
Simone Tebet (MDB)	simone tebet, tebet
Felipe D'avila (Novo)	felipe davila, felipe d'avila

Após cada busca, os *tweets* encontrados foram salvos em um arquivo contendo a data da publicação, id, texto, usuário e região. Ao final da coleta, cada candidato listado possuia um arquivo de *tweets* para cada região do país, totalizando portanto, cinco.

3.2 Criação da base de dados rotulada

Neste trabalho, para classificar cada texto como positivo, negativo ou neutro, utilizou-se os algoritmos *Naive Bayes* e *SVM*. Por serem modelos supervisionados, foi necessário a construção de uma base de *tweets* rotulados, a fim de obter uma base de treinamento para os modelos.

Inicialmente, a fim de analisar a aplicabilidade do conceito de *transfer learning*, o qual segundo Weiss, Khoshgoftaar e Wang (2016) consiste em utilizar um modelo pré-treinado em novos conjuntos de dados, buscou-se bases de dados rotuladas de trabalhos de eleições anteriores. O modelo desenvolvido no estudo realizado por Cristiani, Lieira e Camargo (2020), analisado anteriormente na seção 2.5, foi escolhido para aplicação neste trabalho, devido à facilidade de acesso às bases de dados e códigos utilizados para treinamento dos modelos[4]. A Tabela 3 apresenta os dados contidos na base de treinamento desses autores.

Posteriormente, a fim de comparar o desempenho do modelo construído a partir da base dos autores Cristiani, Lieira e Camargo (2020), foi construído também uma base rotulada manualmente pela autora. Dessa forma, 200 instâncias escolhidas aleatoriamente de cada candidato analisado foram rotuladas manualmente pela autora. Essa quantidade

[3] https://github.com/JustAnotherArchivist/snscrape
[4] https://github.com/andrecristiani/analise-de-sentimentos-eleicoes-2018

Tabela 3 – Sumarização dos dados contidos na base de treinamento do trabalho Cristiani, Lieira e Camargo (2020)

Sentimento	Quantidade de *tweets*
Positivo	250
Neutro	291
Negativo	311
Total	853

foi escolhida devido a proximidade em relação à quantidade de dados utilizada na base rotulada pelos autores Cristiani, Lieira e Camargo (2020).

A Tabela 4 apresenta os dados contidos na base de dados rotulada manualmente pela autora.

Tabela 4 – Sumarização dos dados contidos na base de treinamento rotulada manualmente

Sentimento	Quantidade de *tweets*
Positivo	320
Neutro	297
Negativo	383
Total	1000

Em seguida, as duas bases de dados foram submetidas às técnicas de processamento descritas na seção 3.3.

3.3 Pré-processamento

A etapa de pré-processamento é essencial para o processo de mineração de dados. Neste trabalho, ela foi realizada através da linguagem de programação *Python*[5]. Para a limpeza dos dados, foram realizados os seguintes procedimentos:

- **Remoção de caracteres especiais, hiperlinks, pontuações, marcações de usuários e acentos:** não agregam valor para a construção do modelo;
- **Remoção da palavra utilizada na busca**;
- **Correção de palavras**: palavras escritas de forma incorreta e gírias foram corrigidas[6];
- **Padronização do texto em letras minúsculas:** auxilia na identificação de palavras iguais;

[5] Código disponível em bit.ly/preProcessamento
[6] Dicionário utilizado disponível em bit.ly/dicionarioPython

- **Remoção de stopwords:** utilização da biblioteca NLTK do *Python* para remover palavras que são consideradas irrelevantes para a classificação. Foi criada uma lista com palavras adicionais a fim de completar a biblioteca[7];
- **Tokenização:** extração de unidades mínimas de texto dos *tweets*; e
- **Lematização:** redução das palavras para sua raiz, retirando as inflexões presentes.

Posteriormente, como os dados coletados se tratam de textos, foi necessário aplicar transformações para representá-los em números, a fim de facilitar a construção do modelo.

A codificação escolhida foi o TF-IDF, descrito anteriormente na seção 2.4.3, por se tratar de uma representação que equilibra a importância dos termos frequentes e não-frequentes no texto. Por outro lado, para transformar a polaridade presente nas bases rotuladas (Positivo, Negativo e Neutro), aplicou-se a ferramenta *LabelEncoder*, presente no *Python*, a fim de transformar cada polaridade em um número que a representa. Essa etapa é necessária pois a biblioteca *Sklearn*, a qual foi utilizada para construção do *Naive Bayes* e *SVM*, necessita que a polaridade seja um número. Dessa forma, a polaridade Positivo foi modificada para 2, Neutro para 1 e Negativo para 0.

3.3.1 Representação dos dados

Com a finalidade de validar as técnicas de pré-processamento realizadas, foram construídas nuvens de palavras utilizando a biblioteca *Wordcloud* presente na linguagem de programação *Python*.

Na Figura 5, temos o código desenvolvido para construção de cada nuvem. Os parâmetros passados em relação à cor de fundo e tamanho da imagem foram os padrões, porém é possível personalizar esses aspectos, bem como a cor de cada palavra e formato da nuvem também.

```
# Nuvem de palavras antes da remoção de stopwords
words = df['nuvem_palavras']
all_words = " ".join(w for w in words)
stopwords_ = set(STOPWORDS)

def plot_wordcloud(wc):
    fig, ax = plt.subplots(figsize=(16,8))
    ax.imshow(wc, interpolation='bilinear')
    ax.set_axis_off()
    plt.imshow(wc)

wc = WordCloud(stopwords = stopwords_, background_color='white', width=1600, height=800).generate(all_words)
plot_wordcloud(wc)
```

Figura 5 – Exemplo do código desenvolvido para construção da nuvem de palavras

[7] Lista completa disponível em bit.ly/stopwordsPython

3.4 Construção do modelo

Para a classificação, foram testados inicialmente 2 classificadores: *Naive Bayes* e *SVM*. Eles foram escolhidos por serem os modelos mais utilizados em trabalhos que serviram de referência para este, descritos anteriormente na seção 2.5. É importante destacar que ambos foram construídos na linguaguem de programação *Python*, utilizando para isso a biblioteca *sklearn*.

Na Figura 6 temos um exemplo do código desenvolvido para a construção do modelo.

```
1  from sklearn import naive_bayes, svm
2
3  #Classificador Naive Bayes
4  Naive = naive_bayes.MultinomialNB()
5  Naive.fit(Train_X_Tfidf,Train_Y)
6  predictions_NB = Naive.predict(Test_X_TfidfBolsonaro)
7
8  #Classificador SVM
9  SVM = svm.SVC()
10 SVM.fit(Train_X_Tfidf,Train_Y)
11 predictions_SVM = SVM.predict(Test_X_TfidfBolsonaro)
```

Figura 6 – Exemplo do código desenvolvido para construção dos classificadores

Por fim, foram analisadas as métricas de cada modelo, descritas na seção seguinte.

3.5 Avaliação dos classificadores

Para avaliar o resultado gerado pelo classificador, é necessário a aplicação de métricas de avaliação. Na mineração de dados, a avaliação é realizada por meio de métricas como acurácia, precisão, revocação, *f-measure*, dentre outras. Neste estudo, optou-se por utilizar como parâmetro de avaliação dos modelos a acurácia, precisão e revocação, descritas anteriormente na seção 2.4.5.

É importante destacar que, em relação à base pré-rotulada construída pelos autores Cristiani, Lieira e Camargo (2020), os modelos foram treinados nela e testados na base rotulada manualmente pela autora. Dessa forma, através dos indicadores analisados posteriormente, foi possível analisar o quão bem uma base de eleições anteriores se aplica na atual. Por outro lado, em relação à base rotulada manualmente pela autora, os classificadores foram treinados e testados no mesmo conjunto de dados.

Como estratégia de validação para as duas bases diferentes, foi escolhido o 10-fold cross-validation, o qual consiste na divisão aleatória do conjunto de dados em 10 partes diferentes, sendo que cada parte individual é utilizada como conjunto de teste, e as restantes como conjunto de treinamento.

Na Figura 7, temos o código em *Python* utilizado para avaliar os classificadores.

```
from sklearn.model_selection import cross_validate

nome_metricas = ['accuracy', 'precision_macro', 'recall_macro']
metricas = cross_validate(SVM, X, y, cv=10, scoring=nome_metricas)
print('--------- SVM ---------')
for met in metricas:
    print(f"- {met}:")
    print(f"-- {metricas[met].mean()}")
print('\n')

metricas = cross_validate(Naive, X, y, cv=10, scoring=nome_metricas)
print('--------- Naive Bayes ---------')
for met in metricas:
    print(f"- {met}:")
    print(f"-- {metricas[met].mean()}")
print('\n')
```

Figura 7 – Exemplo do código construído para avaliar os classificadores

3.6 Interpretação dos resultados

Afim de verificar se o objetivo deste estudo foi alcançado, algumas estratégias foram traçadas:

- Criação de gráficos contrastando a aprovação no *Twitter* (sentimento positivo) com as pesquisas eleitorais e o resultado final do primeiro turno;
- Análise da aprovação de cada candidato por região: construção de gráficos para cada região com os sentimentos resultantes da classificação, constrastando com o percentual de votos que cada candidato recebeu naquela região;
- Análise temporal da aprovação de cada candidato, obtida pela classificação dos *tweets*, entre os meses de junho e outubro, contrastando com eventos externos a fim de analisar o impacto que exerceram no aumento ou diminuição da aprovação dos candidatos.

3.7 Considerações finais

Para o desenvolvimento deste estudo, foram utilizadas etapas da Análise de Sentimentos, a fim de extrair informações úteis do *Twitter*, para que o objetivo deste estudo possa ser alcançado.

Na primeira etapa, foi realizada a coleta de dados referentes aos candidatos da eleição de 2022, e logo em seguida, 1000 instâncias foram rotuladas manualmente. Na segunda etapa, foi analisada a aplicabilidade do *transfer learning* na Análise de Sentimentos em eleições, por meio da comparação de classificadores treinados em uma base de dados referente às eleições de 2018, com outros treinados em uma base de dados referente à eleição atual (2022), rotulada manualmente pela autora.

Dessa forma, foi possível testar 2 classificadores para cada base de dados disponível. Cada um foi analisado por meio de métricas previamente escolhidas, e o melhor foi utilizado para classificar os *tweets* da base completa.

Por fim, após a classificação da base completa, foi possível comparar o índice de aprovação de cada candidato nas redes sociais, e seu desempenho real nas eleições, a fim de verificar se há uma correlação entre os dois fatores.

4 Resultados

Este capítulo tem como objetivo analisar os resultados obtidos por cada etapa apresentada anteriormente no capítulo 3, desde a coleta dos dados até a classificação dos *tweets*.

A seção 4.1 sumariza a quantidade de *tweets* coletados por candidato e região, analisando portanto uma possível relação entre esses resultados e o resultado do 1° turno das eleições.

A seção 4.2 e 4.3 apresentam os resultados obtidos com o pré-processamento dos dados, discutindo também a qualidade do pré-processamento realizado.

A seção 4.4 apresenta o resultado dos classificadores que foram testados, por meio de matrizes de confusão e análise de indicadores como acurácia.

A seção 4.5 apresenta os resultados obtidos com o classificador escolhido, *SVM*.

A seção 4.6 contrasta os resultados obtidos com o classificador, com as pesquisas eleitorais realizadas e o resultado final do 1° turno.

4.1 Coleta de dados

A coleta foi realizada na rede social *Twitter* por meio das palavras chaves apresentadas anteriormente na Tabela 2. A data inicial da coleta foi 01 de junho de 2022, e a final 01 de outubro de 2022, 1 dia antes da eleição do 1° turno. No total, foram coletados 6.778.481 *tweets*. A Tabela 5 apresenta a quantidade de *tweets* recuperados para cada candidato.

Tabela 5 – Quantidade de *tweets* coletados por candidato

Candidato	Quantidade de *tweets*
Jair Bolsonaro (PL)	3.028.562
Lula (PT)	2.908.950
Ciro Gomes (PDT)	739.305
Simone Tebet (MDB)	91.892
Felipe d'Avila (Novo)	9.772

A Figura 8 apresenta a quantidade de *tweets* recuperados por candidato em cada mês. É possível perceber um aumento na quantidade de *tweets* no mês de agosto, devido à oficialização da candidatura dos candidatos à presidência, e início da campanha eleitoral. Além disso, nota-se que os dois candidatos mais comentados nas redes, Jair Bolsonaro e Lula, também foram aqueles mais votados e que consequentemente chegaram ao 2° turno.

Por outro lado, Ciro Gomes, 3° colocado no ranking de comentários extraídos na rede social, não obteve um bom resultado nas eleições, ficando atrás de Simone Tebet.

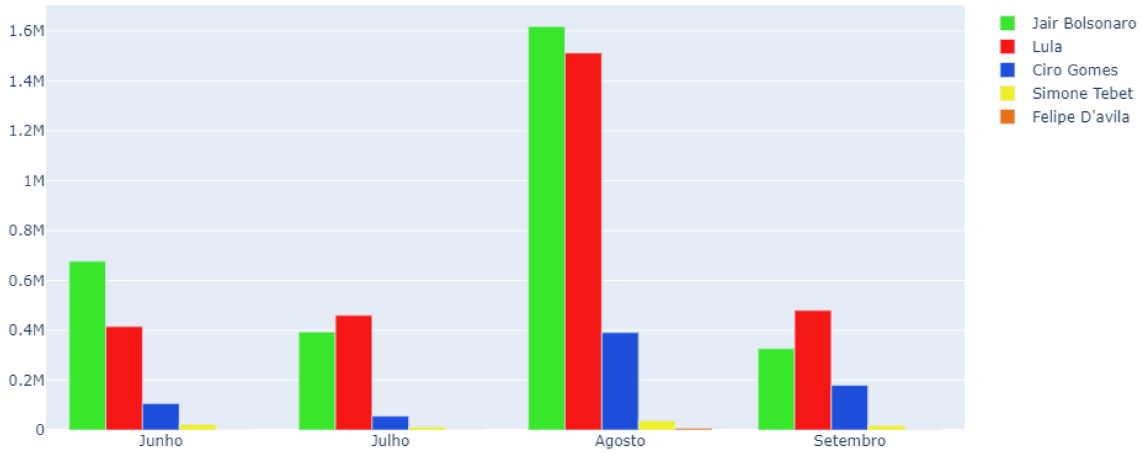

Figura 8 – *Tweets* recuperados por candidato em cada mês

É interessante destacar também o cenário que o *Twitter* se encontrava no dia anterior às eleições. Era esperado que o candidato com mais comentários também fosse aquele que possuiria a maior quantidade de votos, fato que se comprovou como é possível observar na Figura 9. Lula, candidato com mais menções, também foi aquele que mais recebeu votos, ficando em 1°lugar no 1° turno.

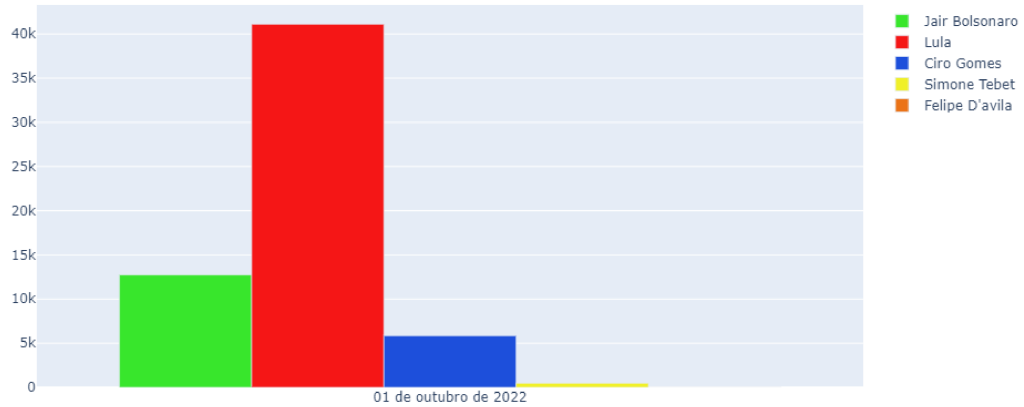

Figura 9 – *Tweets* recuperados por candidato no dia anterior à eleição

Em relação à distribuição dos *tweets* por regiões do Brasil, na Figura 10 é possível ver a quantidade de *tweets* totais por regiões, já a Figura 11 apresenta a distribuição dos *tweets* de cada candidato por região. As regiões que se destacam na quantidade de *tweets* coletados são Sudeste e Centro-Oeste. Sudeste se destaca por ser o maior colégio eleitoral do país, segundo Resende (2022), já a região Centro-Oeste pode-se presumir que seja por causa principalmente de Brasília, centro da política do país. Nesse caso, a má distribuição de *tweets* coletados por região pode gerar enviesamento dos resultados, pois um candidato que se destaca em uma região com mais dados, também se destacará nos resultados como um todo. Dessa forma, é necessário realizar também a classificação e predição por regiões, a fim de diminuir o enviesamento do modelo.

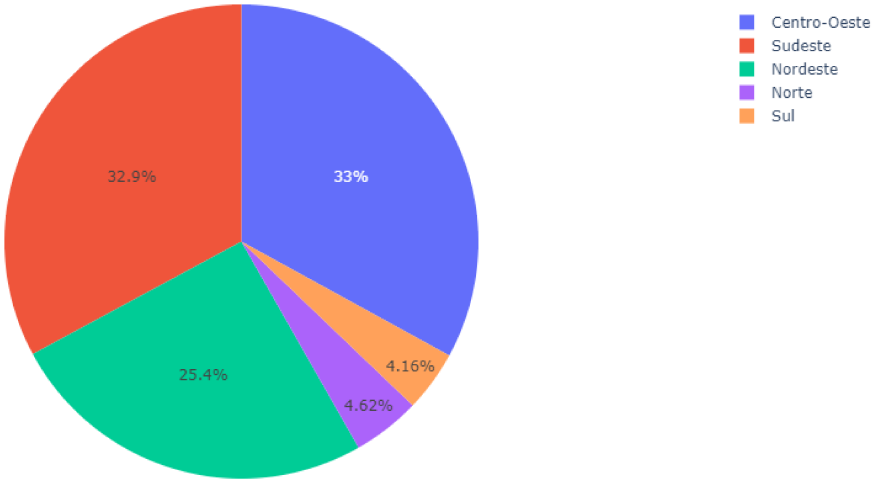

Figura 10 – Distribuição dos *tweets* recuperados por região

Apesar de ser o 3° maior colégio eleitoral do país, a região Sul apresentou uma quantidade maior de *tweets* que a região Norte somente nas menções à Simone Tebet e Felipe d'Avila. Além disso, os candidatos Lula e Felipe d'Avila possuem uma maior quantidade de menções na região Sudeste, ao contrário dos demais candidatos.

Outro destaque importante é a quantidade de menções à Ciro Gomes na região Nordeste, que pode ter sido impulsionada por cargos políticos exercidos anteriormente na região.

4.2 Pré-processamento e representação dos dados

A etapa de pré-processamento é de extrema importância para a Mineração de Textos, pois ela descarta o que não é considerado importante para a construção do mo-

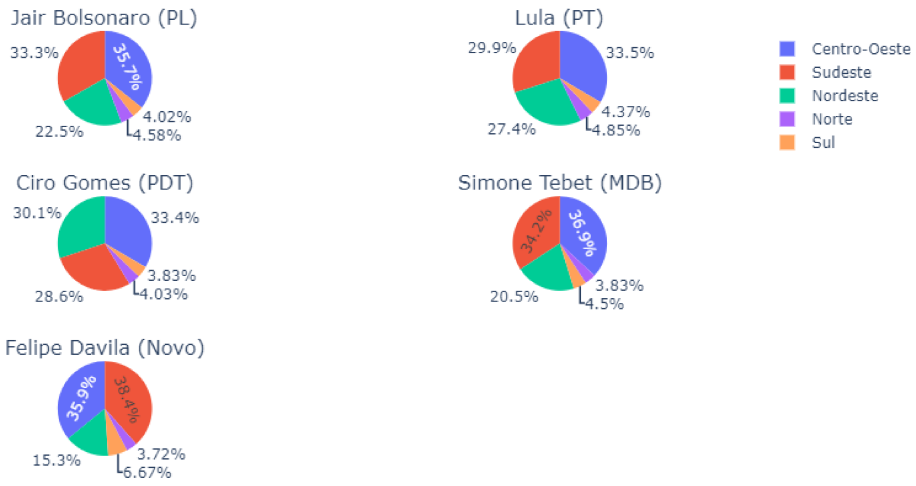

Figura 11 – Distribuição dos *tweets* recuperados por região e por candidato

delo, podendo alterar também o formato dos dados a fim de auxiliar posteriormente na classificação.

Nesta seção é apresentado as etapas realizadas durante esse momento, assim como as nuvens de palavras construídas com a finalidade de representar os dados obtidos na coleta de cada candidato, sendo utilizadas também para validar o pré-processamento realizado.

4.2.1 Pré-processamento

Durante a etapa de pré-processamento, o objetivo principal a ser alcançado é a limpeza e transformação dos dados para que eles possam ser classificados posteriormente. As etapas realizadas e alguns exemplos de resultados obtidos foram:

1. **Remoção de caracteres especiais, *tweets* duplicados, *hiperlinks*, pontuações, marcações de usuários e acentos:** não agregam valor para a construção do modelo. A Tabela 6 apresenta o resultado do pré-processamento de um *tweet* escolhido aleatoriamente ;
2. **Remoção da palavra utilizada na busca**;
3. **Correção de palavras**: palavras escritas de forma incorreta e gírias foram corrigidas. A Tabela 7 apresenta um exemplo de *tweet* antes e após a correção;
4. **Padronização do texto em letras minúsculas:** auxilia na identificação de palavras iguais;

Tabela 6 – Exemplo de um *tweet* antes e após a 1° etapa do pré-processamento

Antes do pré-processamento	**Após o pré-processamento**
@pablomarcal Acho excelente que Bolsonaro e Lula não participem dos debates, já sabemos o que eles vão fazer. Precisamos ouvir novas propostas e conhecer novos candidatos para que possamos fazer uma boa escolha.	Acho excelente que Bolsonaro e Lula nao participem dos debates ja sabemos o que eles vao fazer Precisamos ouvir novas propostas e conhecer novos candidatos para que possamos fazer uma boa escolha

Tabela 7 – Exemplo de um *tweet* antes e após a correção de palavras

Antes do pré-processamento	**Após o pré-processamento**
vc q vai na tebet amanhã pense bem	voce que vai na tebet amanhã pense bem

5. **Remoção de *stopwords*:** utilização da biblioteca NLTK no *Python* para remover palavras que são consideradas irrelevantes para a classificação. A Tabela 8 contém alguns exemplos de *stopwords* que foram removidas;

Tabela 8 – Exemplos de *stopwords* que foram removidas no *script*

de	a	o	que
e	é	do	da
em	um	para	com
não	uma	os	no

6. **Tokenização:** extração de unidades mínimas de texto dos *tweets*. Na Tabela 9 é possível visualizar os resultados obtidos nesta etapa para um *tweet* escolhido aleatoriamente;

Tabela 9 – Exemplo de um texto antes e após a aplicação somente da etapa de tokenização

Antes do pré-processamento	**Após o pré-processamento**
@pablomarcal Acho excelente que Bolsonaro e Lula não participem dos debates, já sabemos o que eles vão fazer. Precisamos ouvir novas propostas e conhecer novos candidatos para que possamos fazer uma boa escolha.	['@', 'pablomarcal', 'Acho, 'excelente', 'que', 'Bolsonaro', 'e', 'Lula', 'não', 'participem', 'dos', 'debates', ',', 'já', 'sabemos', 'o', 'que', 'eles', 'vão', 'fazer', '.', 'Precisamos', 'ouvir', 'novas', 'propostas', 'e', 'conhecer', 'novos', 'candidatos', 'para', 'que', 'possamos', 'fazer', 'uma', 'boa', 'escolha', '.']

7. **Lematização:** redução das palavras para sua raiz, retirando as inflexões presentes. A Tabela 10 apresenta um exemplo de um texto antes e depois da lematização.

Por fim, a Tabela 11 apresenta a quantidade de *tweets* e palavras presentes antes e depois do pré-processsamento. O principal motivo para a redução da quantidade de *tweets* coletados foi a remoção de dados duplicados, que chegaram a compor em torno de 50% da base da maioria dos candidatos. Isso aconteceu devido ao parâmetro *near* da API de coleta, que seleciona os *tweets* feitos próximos à região da cidade informada no

Tabela 10 – Exemplo de um texto antes e após a lemantização

Antes do pré-processamento	**Após o pré-processamento**
@pablomarcal Acho excelente que Bolsonaro e Lula não participem dos debates, já sabemos o que eles vão fazer. Precisamos ouvir novas propostas e conhecer novos candidatos para que possamos fazer uma boa escolha.	['achar', 'excelente', 'bolsonaro', 'Lula', 'participem', 'debate', 'saber', 'vaoer', 'fazer', 'precisar', 'ouvir', 'novo', 'proposta', 'conhecer', 'novo', 'candidato', 'possar', 'fazer', 'bom', 'escolhar']

parâmetro. Como os dados foram coletados utilizando todas as capitais brasileiras, alguns *tweets* foram coletados mais de uma vez pois foram considerados "próximos"a mais de uma cidade. Em relação à remoção de palavras, o principal motivo é a retirada de *stopwords* nas primeiras etapas do pré-processamento.

Tabela 11 – Resumo da base de dados antes e após o pré-processamento

	Quantidade de *tweets*		Quantidade de palavras	
	Antes do pré-processamento	Após o pré-processamento	Antes do pré-processamento	Após o pré-processamento
Total	6.778.481	3.481.203	4.826	3.171

4.2.2 Representação dos dados

Com a finalidade de verificar se as técnicas de pré-processamento utilizadas foram eficazes, foram construídas nuvens de palavras. Além disso, também foi possível analisar as principais discussões que envolviam cada candidato, através das palavras de maior frequência em cada base de dados. É importante destacar porém que, *hashtags* não foram removidas na etapa de pré-processamento, estando portanto presentes também nas nuvens de palavras construídas.

A Figura 12 apresenta as palavras mais comentadas na base do candidato Jair Bolsonaro. É possível perceber algumas palavras em destaque relacionadas à discussões sobre o desempenho de seu primeiro mandato, como reeleger, governo, passar fome, *fake news*, culpa, mito, mau, entre outros. Além disso, percebe-se também que nesse caso, existem mais palavras em tons negativos, se comparadas com outros candidatos.

Outros destaques apresentados na nuvem são jornal nacional, bonner e renata, devido à participações do candidato em debates organizados pelas emissoras de televisão. Por fim, a palavra Lula aparece em grande destaque também, sendo possível concluir que os usuários da rede social fazem comparações frequentes entre os dois candidatos.

A Figura 13 apresenta as palavras mais comentadas na base do candidato Lula. Pode-se observar que contém termos referentes ao tempo que passou na cadeia, como ex-presidiario e ladrão. Também contém termos relacionados a outros candidatos a presidência, como ciro gomes, bolsonaro e bolsonarista. Outro termo bastante comentado foi

Figura 12 – Nuvem de palavras do candidato Jair Bolsonaro (PL)

a palavra bozo, apelido popular utilizado na Internet por quem é contrário ao candidato Jair Bolsonaro. A alta frequência desse termo juntamente com as menções ao candidato Lula, pode indicar a frequente comparação entre os candidatos. Por último, é possível perceber alguns termos positivos, como bom, amigo, amar, apoiar e eleger.

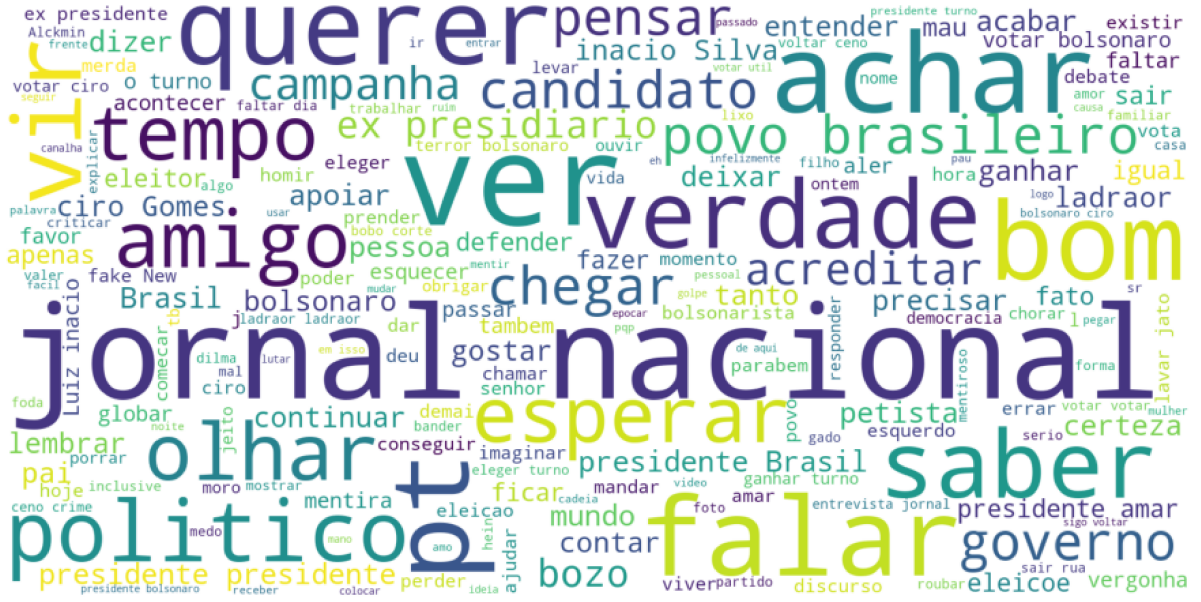

Figura 13 – Nuvem de palavras do candidato Lula (PT)

A Figura 14 apresenta as palavras mais comentadas na base do candidato Ciro Gomes. É possível perceber o resultado da lematização, por meio de palavras como votar, falar, fazer, dizer, que estão sem inflexões. Uma das principais discussões envolvendo Ciro Gomes é a possibilidade de ele chegar ao segundo turno sendo um candidato de terceira via,

ou seja, uma opção diferente para aqueles que não querem votar em Lula e Bolsonaro, os dois candidatos com maior intenção de voto das eleições de 2022. Inclusive, palavras como petista e bolsonarista complementam essa possibilidade. Outras palavras em destaque para o candidato são entrevista e jornal nacional, referente as suas participações nos debates, que geralmente são elogiadas nas redes sociais devido à sua boa habilidade de comunicação. Além disso, é possível observar também que existem palavras de cunho positivo como bom, gostar, ganhar, votar e amar indicando possíveis elogios ao candidato. Por outo lado, palavras de cunho negativo também estão presentes como mau e perder.

Figura 14 – Nuvem de palavras do candidato Ciro Gomes (PDT)

A Figura 15 apresenta as palavras mais comentadas na base de Simone Tebet. Dentre elas, temos termos relacionados à participação da candidata na CPI do Covid-19, como CPI, Renan Calheiros, corrupto, dinheiro público, covid e circo (expressão que ficou conhecida nas redes sociais para se referir à CPI). Outros termos de destaque se referem a uma possível ida para o segundo turno, já que assim como Ciro Gomes (termo que também aparece em destaque), foi uma das opções mais fortes da terceira via. Além disso, um dos maiores destaques é a palavra mulher, que pode ser uma referência à chapa totalmente feminina da candidata.

Por último, a Figura 16 apresenta as palavras mais comentadas na base de Felipe d'Avila. Como destaque, temos bosopetismo, termo popularmente conhecido na internet para se referir as similaridades de discursos políticos do bolsonarismo e petismo, como distorção de fatos, negacionismo político e ausência de autocrítica (MORI, 2020). Outro termo possivelmente relacionado à esse é acorda brasil, expressão bastante utilizada nas redes sociais para se referir ao movimento da terceira via, ou seja, a busca por outro candidato que não seja Lula ou Jair Bolsonaro, que inclusive também foram altamente

Figura 15 – Nuvem de palavras da candidata Simone Tebet (MDB)

mencionados juntamente com o candidato em questão.

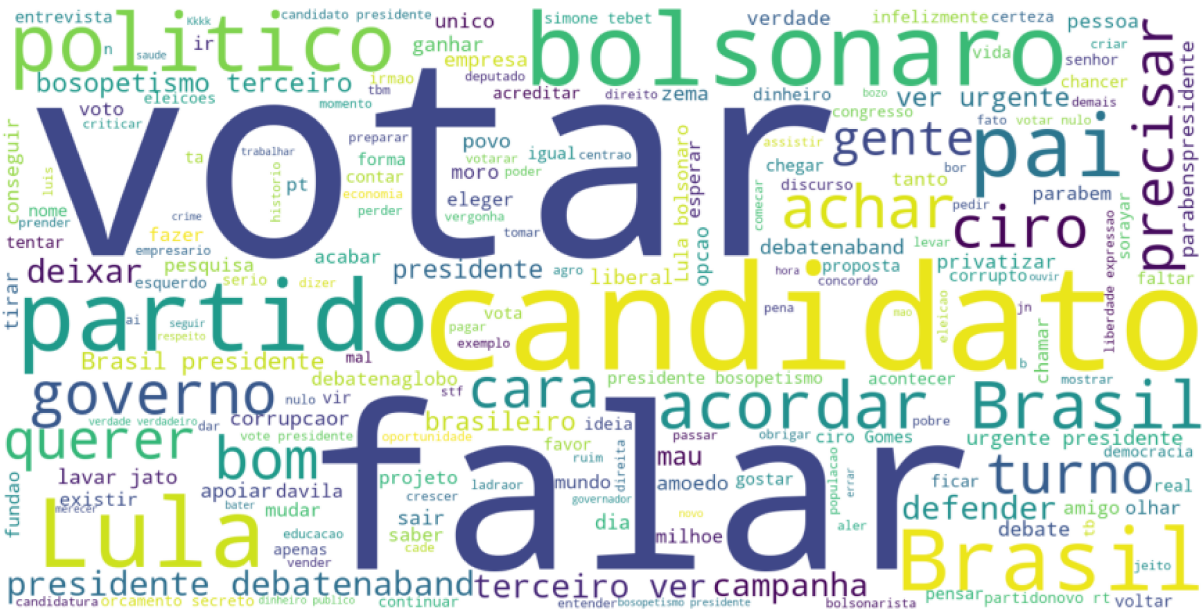

Figura 16 – Nuvem de palavras do candidato Felipe d'Avila (Novo)

4.3 Avaliação do classificador

Como descrito anteriormente nas seções 3.2 e 3.4, foram testados 2 algoritmos de classificação (*Naive Bayes* e *SVM*) para cada base de dados de treinamento. Com o objetivo de analisar qual base de treino e qual classificador foi o melhor dentre as opções

descritas, as medidas de avaliação acurácia, precisão e revocação, mencionadas na seção 3.5, foram analisadas.

Para os dois experimentos, foi utilizado o método *10-fold-cross-validation*, definido anteriormente na seção 2.4.5. Portanto, para a geração da matriz de confusão e indicadores, a base de dados foi dividida em treino e teste 10 vezes, de forma que cada instância fizesse parte da base de treino 9 vezes, e teste apenas 1 vez. Para a geração da matriz de confusão, os resultados apresentados na tabela são a soma das polaridades obtidas a cada partição realizada. Por outro lado, os indicadores de acurácia, precisão e revocação representam a média dos valores obtidos em cada uma das partições.

4.3.1 Experimento 1

O primeiro experimento realizado construiu o modelo de decisão a partir de uma base de dados obtida da Internet e construída pelos autores Cristiani, Lieira e Camargo (2020), referente a dados do *Twitter* das eleições de 2018. A base contém 853 instâncias referentes aos candidatos da eleição presidencial de 2018, e na Tabela 3 é possível ver a quantidade de *tweets* para cada polarização. Os modelos construídos pelos autores do trabalho (CRISTIANI; LIEIRA; CAMARGO, 2020) foram treinados nos algoritmos *Naive Bayes* e *SVM*, sendo o último o escolhido para gerar a classificação, pois obteve o melhor resultado - 66,66% de acurácia.

O objetivo do experimento foi verificar o quão bem o *transfer learning* se aplica neste estudo. Assim, um modelo de decisão é construído usando a base de dados rotulada para as eleições de 2018 do trabalho de Cristiani, Lieira e Camargo (2020) e usada para avaliar o sentimento dos *tweets* da base coletada neste trabalho. Para isso, 2 classificadores, sendo eles *SVM* e *Naive Bayes*, foram treinados e tiveram seus resultados comparados.

A base de teste utilizada neste experimento foi rotulada manualmente pela autora. Para isso, foram escolhidas 200 instâncias aleatórias da base de dados de cada candidato, totalizando 1000[1] instâncias portanto. A Tabela 4, presente na seção 3.2, apresenta a quantidade de dados rotulados para cada polarização existente.

A Tabela 12 mostra a matriz de confusão obtida para esse cenário com o *SVM*. É possível observar as principais falhas desse modelo construído: prever as classes positivas, já que 41% dos *tweets* positivos foram classificados como negativos, assim como 54% dos *tweets* negativos foram classificados como neutros. No geral, o classificador acertou 45% das instâncias, um valor baixo.

Por outro lado, a Tabela 13 mostra a matriz de confusão obtida para esse cenário com o *Naive Bayes*. Assim como o *SVM*, porém mais intensificado, o *Naive Bayes*

[1] Após o pré-processamento, a base de dados passou a ter 997 instâncias devido à remoção de dados duplicados.

Tabela 12 – Matriz de confusão para o *SVM* no Experimento 1

		Classificado como			
		Positivo	Neutro	Negativo	Total
Real	Positivo	96 (30%)	91 (29%)	132 (41%)	319
	Neutro	34 (12%)	102 (34%)	160 (54%)	296
	Negativo	47 (12%)	83 (22%)	252 (66%)	382
Total		177	276	544	997

classificou mais classes neutras como negativas. Além disso, ele não conseguiu prever a classe alvo, pois a maioria das instâncias foram classificadas como neutras ou negativas. No geral, o classificador acertou 42% das instâncias.

Tabela 13 – Matriz de confusão para o *Naive Bayes* no Experimento 1

		Classificado como			
		Positivo	Neutro	Negativo	Total
Real	Positivo	100 (32%)	110 (34%)	109 (34%)	319
	Neutro	41 (13%)	126 (43%)	129 (44%)	296
	Negativo	57 (15%)	131 (34%)	194 (51%)	382
Total		198	367	432	997

Além da matriz de confusão, também foi avaliado a acurácia, precisão e revocação de cada classificador, com os resultados descritos na Tabela 14. É possível observar que nos dois modelos, o valor da revocação é o menor em relação as 3 métricas, indicando que eles não conseguem classificar corretamente a classe alvo (positiva).

Tabela 14 – Avaliação dos indicadores Experimento 1

	Naive Bayes	SVM
Acurácia	42%	45%
Precisão	41%	43%
Revocação	43%	45%

Portanto, é possível observar que os resultados obtidos neste experimento não foram bons.

4.3.2 Experimento 2

Com o intuito de melhorar o classificador, foi utilizada a base de teste do experimento anterior, que possui 200 instâncias de cada candidato rotuladas manualmente pela autora, totalizando portanto 1000 instâncias. A Tabela 4, presente na seção 3.2, apresenta a quantidade de dados rotulados para cada polarização existente. Nesta base, também foram treinados e testados os classificadores *SVM* e *Naive Bayes*.

É importante destacar que para a construção desse experimento, a base foi balanceada utilizando o *undersampling*, técnica que consiste na diminuição de instâncias das

classes maioritárias. Dessa forma, a base de dados que antes continha 1000 instâncias, passou a ter 890, sendo 297 instâncias para cada polaridade.

A Tabela 15 mostra a matriz de confusão obtida para o modelo construído com o classificador *SVM*. É possível perceber que a falha principal do modelo foi o reconhecimento errôneo de classes neutras como negativas. Porém, no geral o classificador obteve 52% de acerto, valor maior que o observado nos classificadores do experimento 1.

Tabela 15 – Matriz de confusão para o *SVM* no Experimento 2

		Classificado como			
		Positivo	Neutro	Negativo	Total
Real	Positivo	131 (44%)	76 (26%)	90 (30%)	297
	Neutro	44 (14%)	126 (43%)	126 (43%)	296
	Negativo	29 (10%)	71 (24%)	197 (66%)	297
Total		204	273	413	890

Já a Tabela 16 mostra a matriz de confusão obtida para o modelo construído com o classificador *Naive Bayes*. Nesse caso, é possível perceber que o modelo construído possui dificuldade em aprender classes neutras, já que não conseguiu prever a maioria das instâncias dessa classe corretamente. Por esse motivo, o modelo preveu 50% das instâncias corretamente, número também maior que os classificadores do experimento 1.

Tabela 16 – Matriz de confusão para o *Naive Bayes* no Experimento 2

		Classificado como			
		Positivo	Neutro	Negativo	Total
Real	Positivo	192 (65%)	55 (19%)	50 (16%)	297
	Neutro	112 (38%)	92 (31%)	92 (31%)	296
	Negativo	90 (30%)	46 (15%)	161 (55%)	297
Total		394	193	303	890

Além da matriz de confusão, e assim como para o outro modelo, também foram avaliadas as medidas acurácia, precisão e revocação de cada classificador, com os resultados descritos na Tabela 17. É possível observar que nesse caso todos os indicadores melhoraram para os dois classificadores, com o *SVM* se destacando em todas as métricas.

Tabela 17 – Avaliação dos indicadores Experimento 2

	Naive Bayes	SVM
Acurácia	50%	52%
Precisão	51%	54%
Revocação	50%	51%

Portanto, com base nos dados obtidos na Tabela 17, o modelo escolhido para este projeto foi o construído com o classificador *SVM* do Experimento 2[2], por possuir as

[2] Ao aplicar as mesmas técnicas utilizadas na construção do modelo do Experimento 2, na base dos autores Cristiani, Lieira e Camargo (2020), a acurácia obtida foi de 68%, valor similar ao encontrado por eles na construção de seu projeto.

maiores métricas se comparado com o *Naive Bayes* e o Experimento 1.

4.4 Resultados da predição realizada com o classificador *SVM*

Após a escolha e treinamento do classificador, foi possível classificar toda a base de dados de cada candidato, e assim analisar a popularidade de cada um no *Twitter*. É importante destacar que para as análises de aprovação dos candidatos, os *tweets* classificados como neutros foram retirados, pois não agregam na distinção do que é apoio ou rejeição aos candidatos.

A Figura 17 mostra a visão geral da classificação dos *tweets* dos candidatos. É possível perceber o destaque que os candidatos Ciro Gomes e Felipe d'Avila tiveram nas redes sociais, obtendo as maiores taxas de aprovação dos candidatos analisados. Por outro lado, Simone Tebet, Jair Bolsonaro e Lula tiveram respectivamente as piores taxas de aprovação.

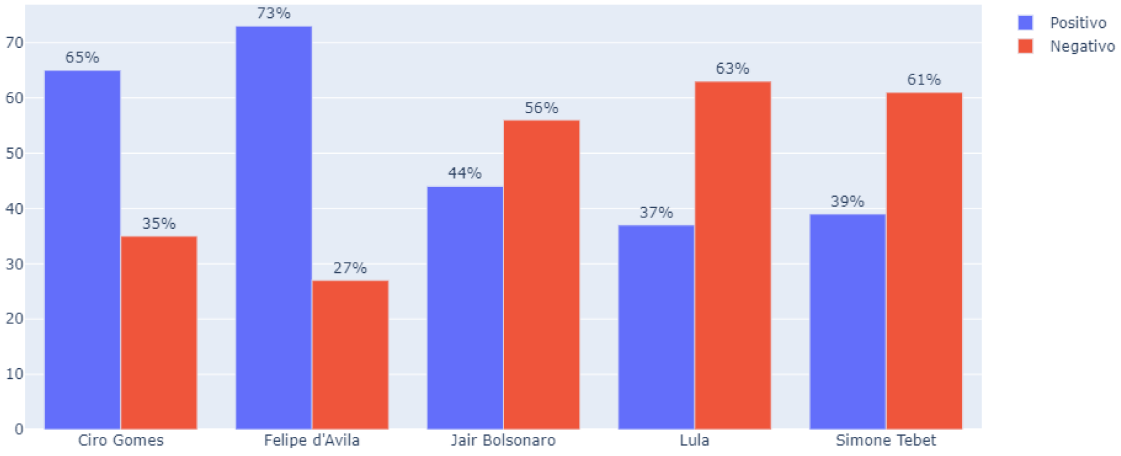

Figura 17 – Porcentagem de aprovação total por candidato

Se o primeiro turno da eleição presidencial de 2022 fosse definido de acordo com a taxa de aprovação dos candidatos na rede social *Twitter*, teríamos uma disputa no segundo turno entre Ciro Gomes e Felipe d'Avila.

4.4.1 Ciro Gomes

Em relação ao candidato Ciro Gomes, podemos analisar nas Figuras 18 e 19 que o candidato possui uma alta taxa de aprovação entre os usuários do *Twitter*. Em todos os meses e regiões do país, o candidato obteve mais comentários positivos do que negativos,

um bom indicador de que talvez o candidato apresente uma ótima participação na eleição presidencial. Além disso, analisando somente a Figura 18, é possível perceber que a taxa de aprovação do candidato foi aumentando ao longo dos meses, com exceção de setembro. Nesse mês, os debates aconteceram com maior frequência e nas principais emissoras de televisão, aumentando portanto as discussões e críticas acerca do candidato.

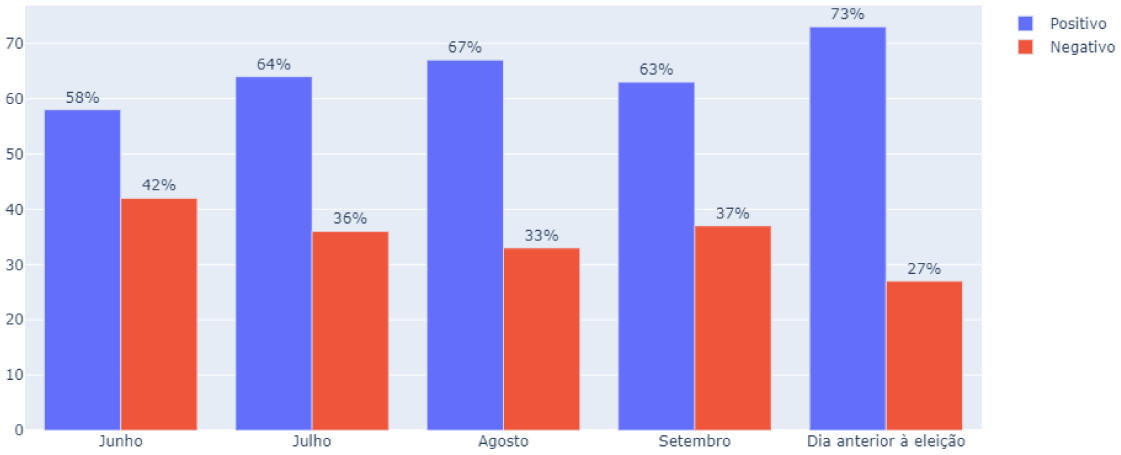

Figura 18 – Aprovação e rejeição ao longo dos meses - Ciro Gomes (PDT)

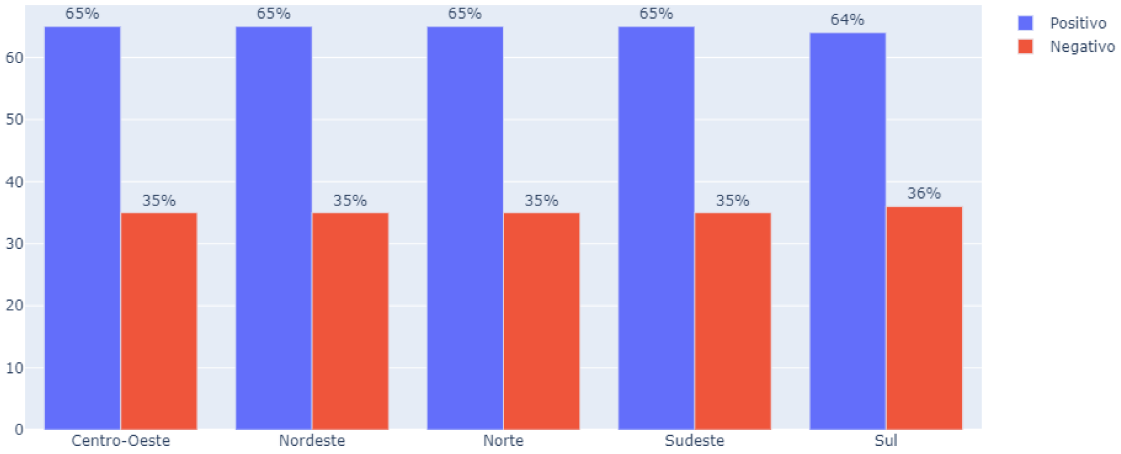

Figura 19 – Aprovação e rejeição por região - Ciro Gomes (PDT)

4.4.2 Felipe d'Avila

O candidato Felipe d'Avila apresentou uma boa aprovação nas redes sociais, fato que pode ser visto na Figura 20. Nota-se que possui apenas 1 mês em que sua taxa de reprovação foi maior que a de aprovação, contudo é importante comentar que o mês de agosto foi o mês no qual os candidatos foram anunciados oficialmente e os debates foram iniciados. Por esse motivo, foi também o pico de menções do candidato, o que pode ter gerado discussões de usuários que não o conheciam anteriormente e por isso sua taxa de reprovação aumentou.

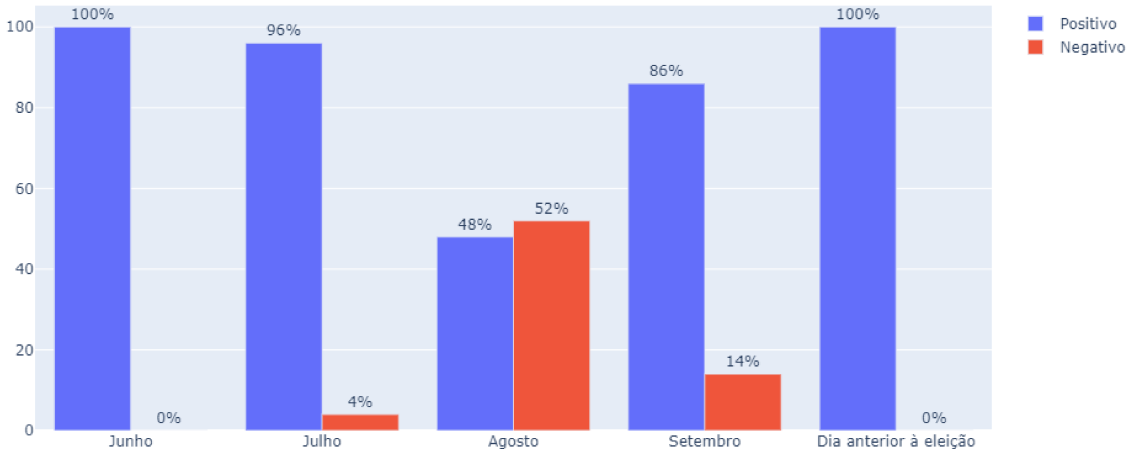

Figura 20 – Aprovação e rejeição ao longo dos meses - Felipe d'Avila (Novo)

A Figura 21 apresenta a quantidade de *tweets* classificados como positivo e negativo por região do Brasil. As regiões com menos dados colhidos para o candidato foram as regiões Norte, Nordeste e Sul, com 1%, 5,2% e 5,5% respectivamente, porém foram justamente essas regiões com mais comentários de aprovação.

Por fim, a taxa de aprovação do candidato no dia antes da eleição estava em 100%. Apesar de não conter comentários negativos, a quantidade de *tweets* é muito baixa, sendo 3 no total, e portanto não é possível afirmar que o candidato terá uma boa participação na eleição presidencial de 2022.

4.4.3 Jair Bolsonaro

O candidato Jair Bolsonaro, como visto anteriormente, apresentou uma taxa de rejeição geral maior que a de aprovação. Por outro lado, a Figura 22 demonstra que, a diferença entre a taxa de aprovação e reprovação diminuiu consideravelmente em se-

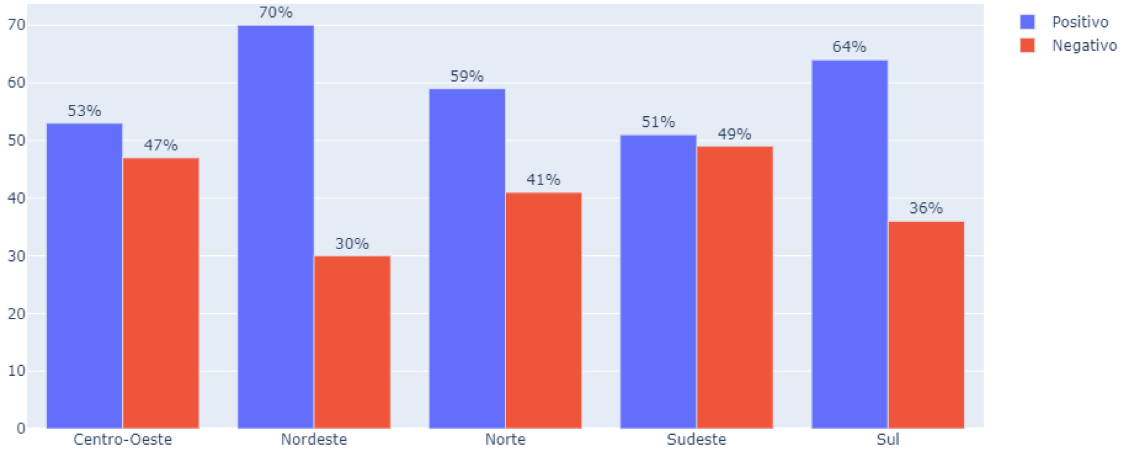

Figura 21 – Aprovação e rejeição por região - Felipe d'Avila (Novo)

tembro, o mês anterior à eleição, indicando uma possível melhora no seu desempenho na eleição presidencial. Além disso, o dia anterior a eleição foi a primeira vez que o candidato apresentou uma taxa de aprovação maior que a de rejeição.

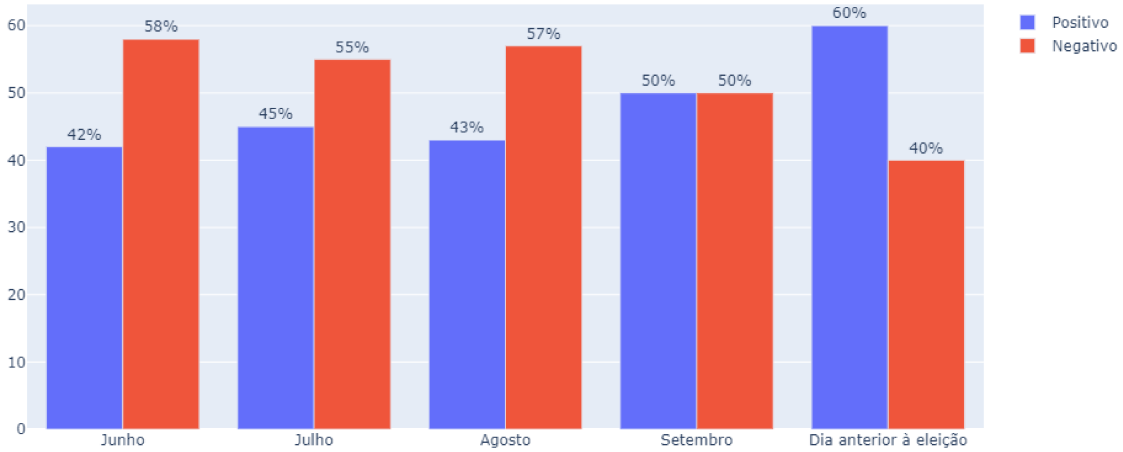

Figura 22 – Aprovação e rejeição ao longo dos meses - Jair Bolsonaro (PL)

A Figura 23 apresenta os dados por região. Nota-se que o Nordeste é a região com a maior taxa de reprovação do candidato. Por outro lado, a região Norte aparece com a maior taxa de aprovação, porém é a região com menos dados disponíveis, representando

apenas 1.8% dos *tweets* do candidato.

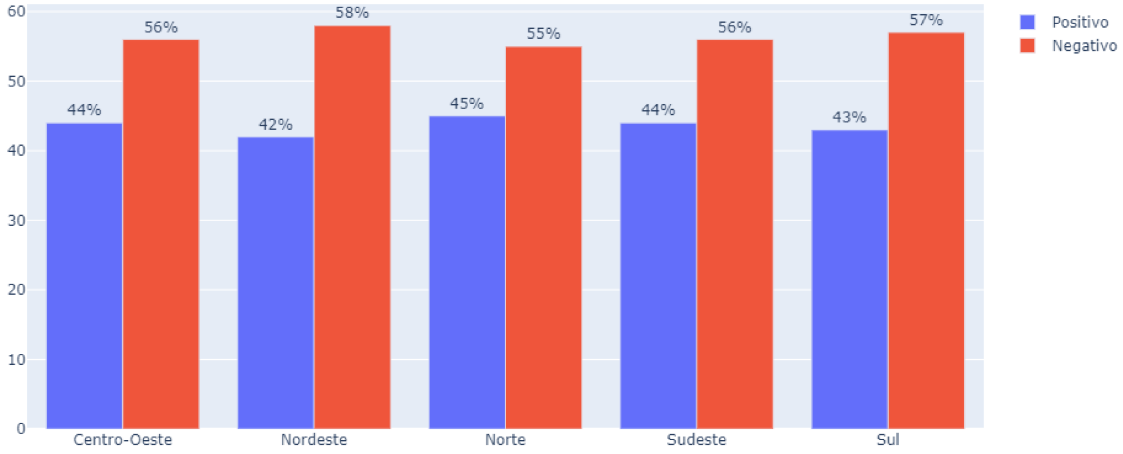

Figura 23 – Aprovação e rejeição por região - Jair Bolsonaro (PL)

4.4.4 Lula

Em relação ao candidato Lula, nota-se na Figura 24 a constante evolução em sua popularidade na rede social analisada, chegando no pico no dia anterior à eleição, indicando que talvez o candidato se destaque na votação. Além disso, assim como candidato Jair Bolsonaro, o dia anterior a eleição foi a primeira vez que o candidato Lula apresentou uma taxa de aprovação maior que a de rejeição.

Ao analisar as regiões de acordo com a Figura 25, percebe-se que a região Nordeste foi a que o candidato mais recebeu comentários positivos na rede social, apesar de ser a 3° região com maior quantidade de comentários do candidato, compondo 10% dos *tweets* totais de sua base de dados. Por outro lado, a região Centro-Oeste, que compõe 58% dos *tweets* do candidato, foi a 3° região com mais comentários negativos, o que pode abaixar sua taxa de aprovação total, justamente por conter mais comentários que qualquer outra região.

4.4.5 Simone Tebet

Em relação à candidata Simone Tebet, percebe-se na Figura 17, que a taxa de rejeição na rede social *Twitter* é bem maior que a de aprovação. Apesar disso, ao analisar somente a Figura 26, nota-se que a taxa de aprovação da candidata evoluiu constantemente ao longo dos meses, chegando ao maior valor no dia anterior à eleição. Talvez essa evolução

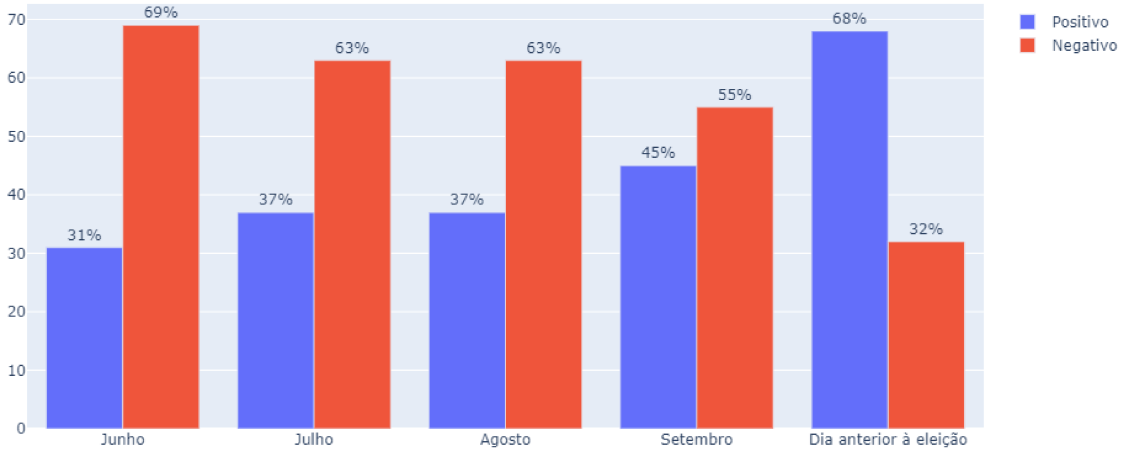

Figura 24 – Aprovação e rejeição ao longo dos meses - Lula (PT)

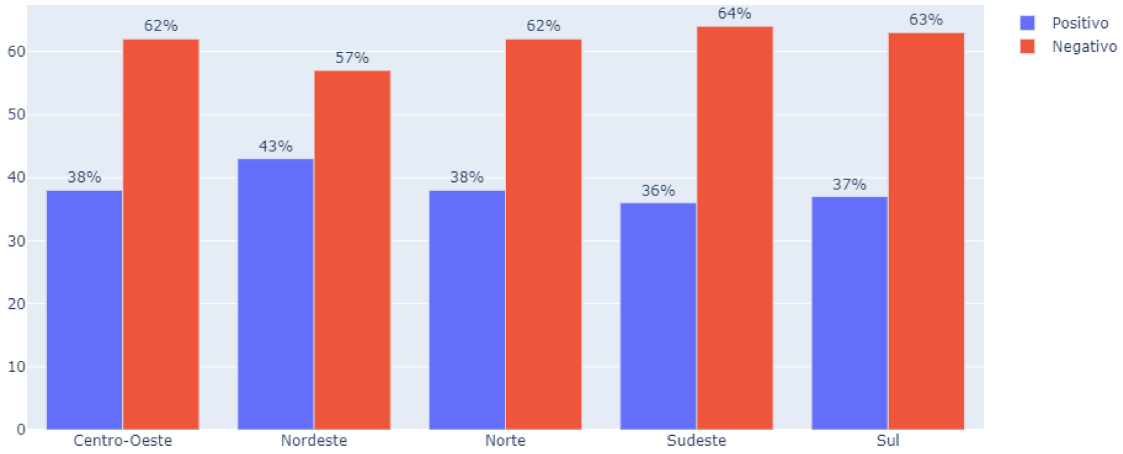

Figura 25 – Aprovação e rejeição por região - Lula (PT)

possa ser um indicativo de que a candidata supreenderá na quantidade de votos totais recebidos.

Ao analisar as regiões de acordo com a Figura 27, nota-se que a região Norte possui a maior taxa de aprovação, porém, compõe apenas 1,4% dos *tweets* da candidata. Além disso, a região Sul, que possui a maior taxa de rejeição, também está sub-representada na base, compondo 4% dos *tweets* totais.

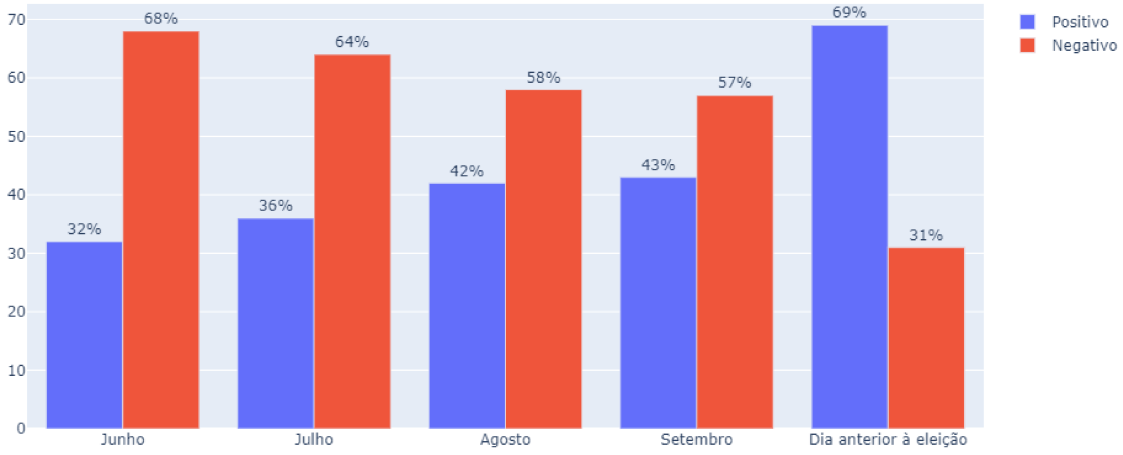

Figura 26 – Aprovação e rejeição ao longo dos meses - Simone Tebet (MDB)

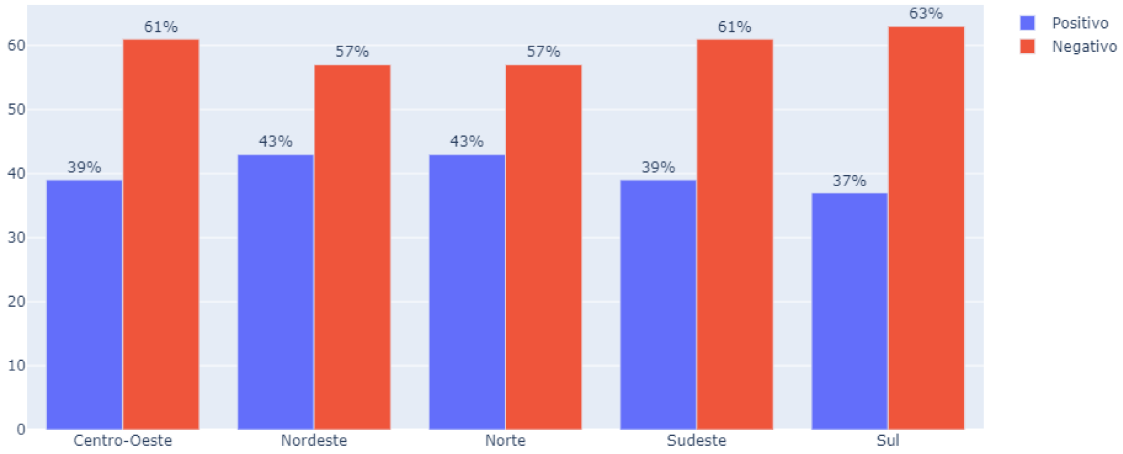

Figura 27 – Aprovação e rejeição por região - Simone Tebet (MDB)

4.5 Contraste dos resultados da classificação e resultados oficiais da eleição presidencial de 2022

Como dito anteriormente, se o resultado do primeiro turno fosse baseado na popularidade dos candidatos nas redes sociais, medida pelo proporção de *tweets* positivos de um candidato, teríamos uma disputa entre Ciro Gomes e Felipe d'Avila. Porém, como mostrado na Figura 28, o resultado oficial do primeiro turno foi diferente do gerado na

classificação dos *tweets*, pois os candidatos classificados para o segundo turno foram Jair Bolsonaro e Lula. Além disso, o candidato que mais recebeu votos foi o menos aprovado na rede social.

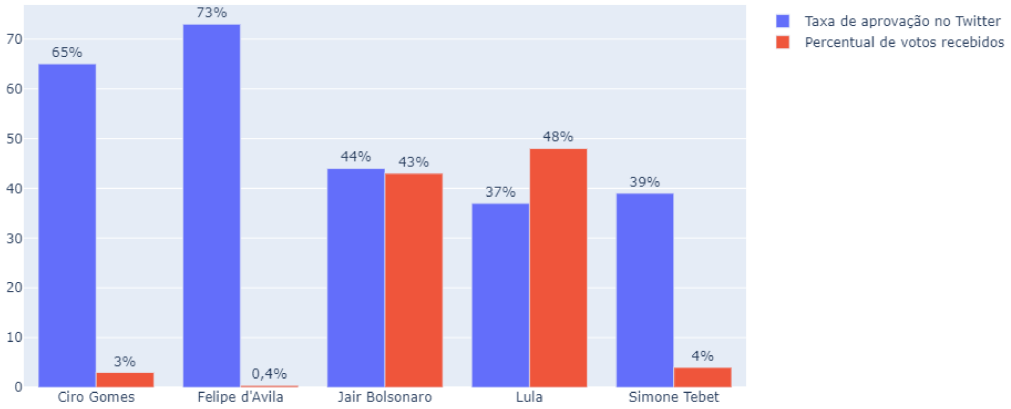

Figura 28 – Comparação entre a aprovação no *Twitter* e o percentual de votos recebidos no 1° turno

Baseado no experimento realizado, foi possível concluir que não houve semelhanças entre a taxa de popularidade do candidato na rede social *Twitter* e seu desempenho na eleição presidencial, com exceção da taxa de aprovação do candidato Jair Bolsonaro se aproximar bastante de seu percentual de votos. Porém, os dois candidatos mais comentados no *Twitter* foram os que conseguiram avançar para o segundo turno.

Em relação à região Sudeste, Jair Bolsonaro obteve a maior quantidade de votos no primeiro turno, porém aparece como terceiro no ranking de popularidade construído neste estudo. O candidato Ciro Gomes, que obteve a maior taxa de aprovação na rede social, ficou em quarto lugar, na frente apenas de Felipe d'Avila na votação presidencial, como mostrado na Figura 29.

No gráfico da Figura 30, retratando os resultados da região Nordeste, é possível perceber que assim como no Sudeste, o candidato Felipe d'Avila, mais bem avaliado nas redes sociais, ficou em último lugar na quantidade de votos recebidos. Além disso, o candidato mais votado na região, Lula, ficou em 3° lugar - empatado com Simone Tebet - no ranking de aprovação na rede social da região.

Em relação à região Sul, analisando a Figura 31, é possível perceber que o terceiro candidato mais bem avaliado no *Twitter*, Jair Bolsonaro, foi o que mais recebeu votos na região. Além disso, os dois candidatos mais bem avaliados ficaram nas 2 últimas posições em relação ao percentual de votos recebidos. Em contrapartida, a taxa de aprovação e o percentual de votos recebidos do candidato Lula ficaram bem próximos.

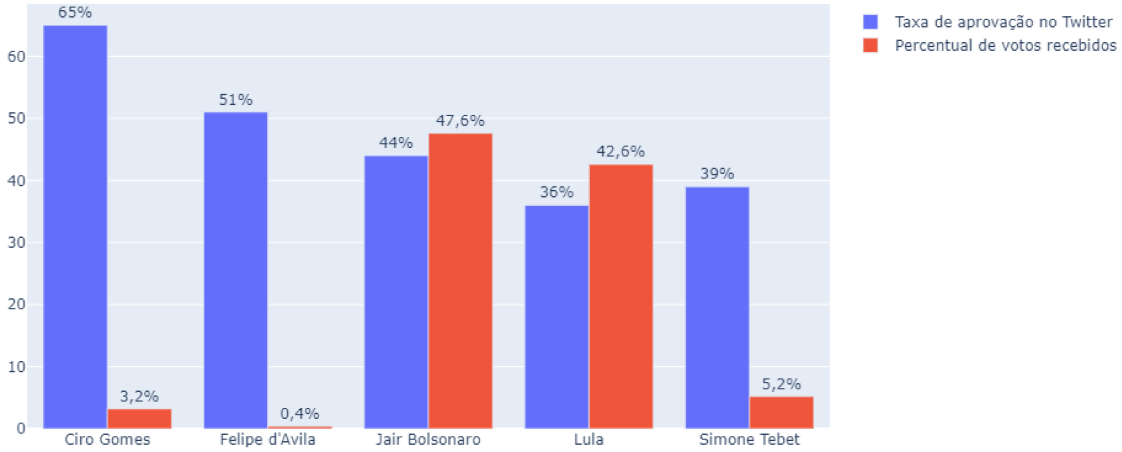

Figura 29 – Comparação entre a aprovação no *Twitter* e o percentual de votos recebidos no 1° turno na região Sudeste

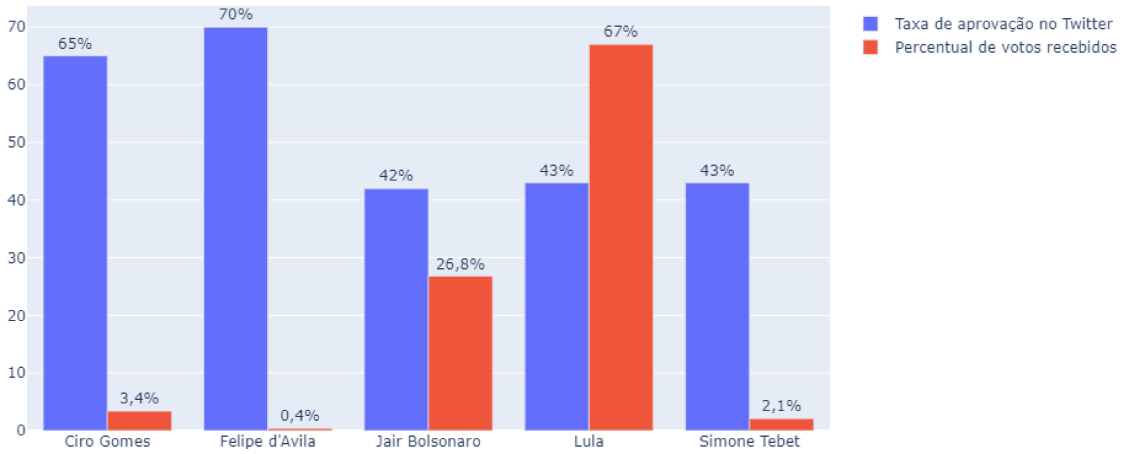

Figura 30 – Comparação entre a aprovação no *Twitter* e o percentual de votos recebidos no 1° turno na região Nordeste

A Figura 32 apresenta a comparação da taxa de aprovação e percentual de votos recebidos na região Norte. É possível perceber que a taxa de aprovação na rede social e o percentual de votos recebidos do candidato Jair Bolsonaro ficaram bem perto. Apesar disso o candidato que obteve a segunda maior quantidade de votos, ficou em terceiro lugar no ranking de popularidade na rede social na região, sendo que o mais votado ficou em

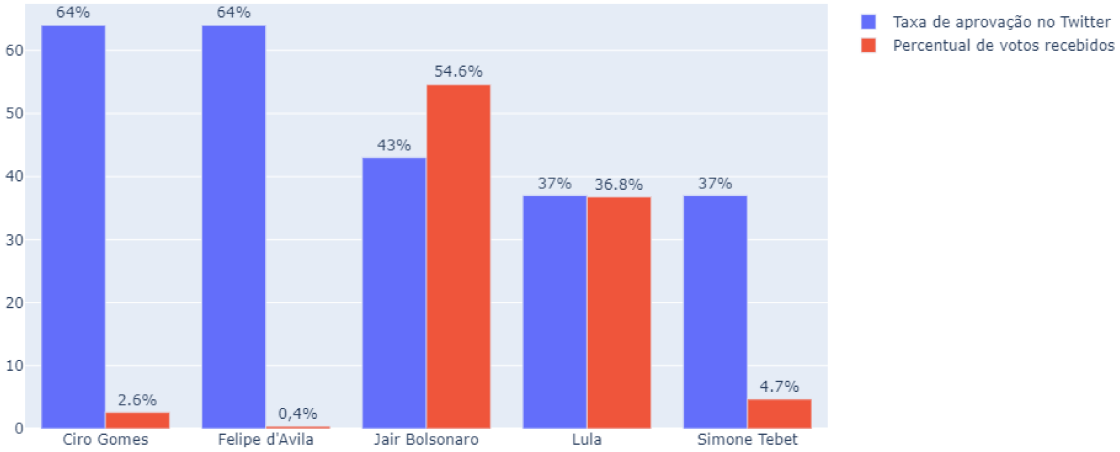

Figura 31 – Comparação entre a aprovação no *Twitter* e o percentual de votos recebidos no 1° turno na região Sul

último no ranking de popularidade. Além disso, assim como observado na região sul, os dois primeiros candidatos mais bem avaliados no *Twitter* ficaram nos dois últimos lugares em relação ao percentual de votos recebidos.

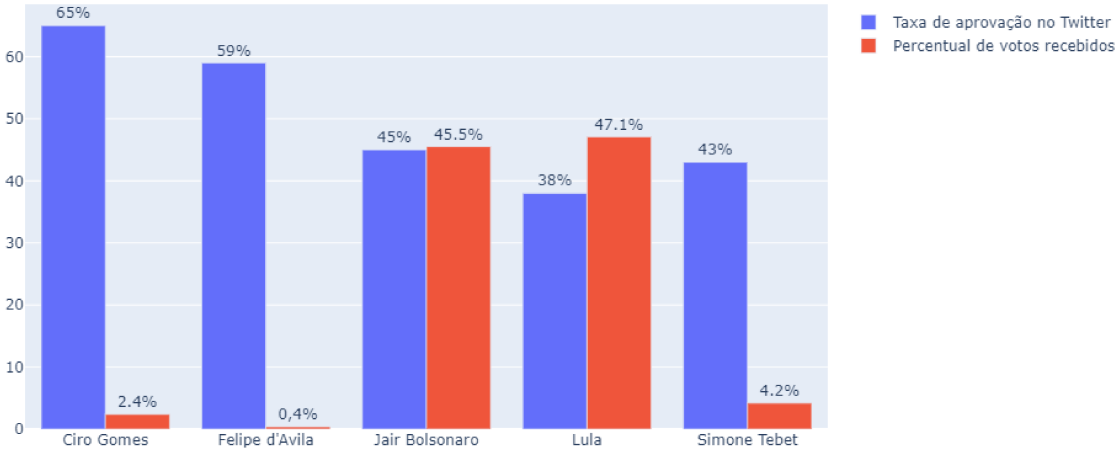

Figura 32 – Comparação entre a aprovação no *Twitter* e o percentual de votos recebidos no 1° turno na região Norte

Por último, em relação à região Centro-Oeste, não houve tamanha semelhança entre os resultados obtidos pelo classificador e o resultado oficial do primeiro turno, como

demonstrado na Figura 33. Uma exceção seria os resultados do candidato Lula, pois sua taxa de aprovação e percentual de votos recebidos ficaram bem próximos.

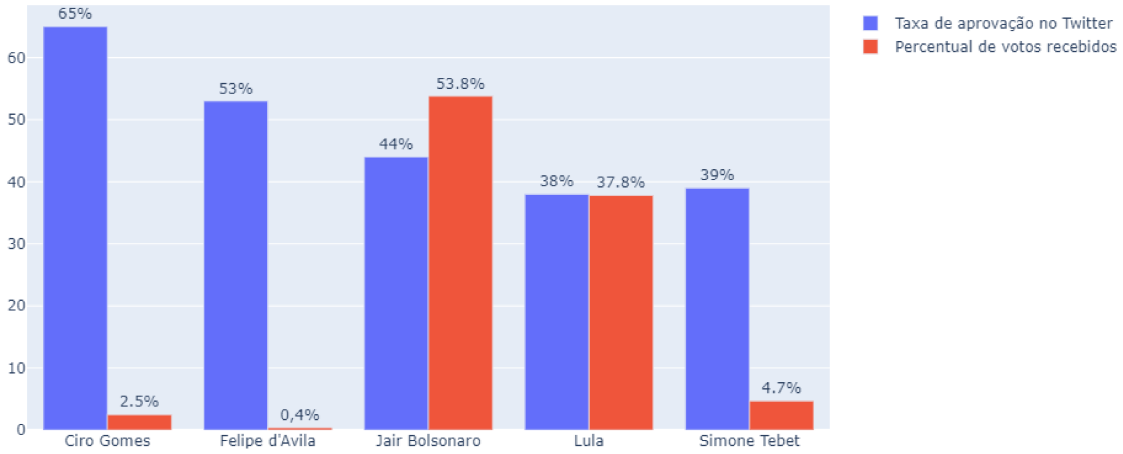

Figura 33 – Comparação entre a aprovação no *Twitter* e o percentual de votos recebidos no 1° turno na região Centro-Oeste

4.5.1 Pesquisas eleitorais

Para realização da comparação com algumas pesquisas eleitorais realizadas durante o período eleitoral, foram escolhidas aquelas realizadas pelo Instituto de Pesquisa e Consultoria Estratégica (IPEC). Os dados das pesquisas analisadas estão disponíveis no site oficial da instituição[3]. Ao todo, três pesquisas em meses diferentes foram analisadas: a primeira referente ao mês de agosto, com dados coletados entre os dias 26 e 28; a segunda referente à setembro, com dados coletados entre os dias 17 e 18; e por último, foi analisada a última pesquisa realizada antes da eleição, com dados entre 29 de setembro e 01 de outubro.

A Figura 34 apresenta os resultados obtidos com as pesquisas eleitorais realizadas. É possível notar algumas semelhanças com a classificação realizada neste estudo. A popularidade da candidata Simone Tebet foi aprimorando ao longo dos meses, fato que se nota perceptível também nas pesquisas realizadas, apesar do crescimento não ter sido tão expressivo quanto o mostrado neste estudo.

Em relação ao candidato Jair Bolsonaro, o aumento da taxa de aprovação observado nas redes sociais, foi maior que o apresentado nas pesquisas eleitorais, apesar de no

[3] https://www.ipec-inteligencia.com.br/pesquisas/

resultado final, o candidato ter surpreendido na quantidade de votos recebidos, um valor bem acima do projetado na maioria das pesquisas realizadas.

No que se refere ao candidato Lula, e assim como o candidato Jair Bolsonaro, o modelo previu o aumento na intenção de votos do candidato, apesar de esse aumento ter sido menos expressivo que o observado na sua taxa de aprovação nas redes sociais. Por outro lado, os resultados dos candidatos Ciro Gomes e Felipe d'Avila divergiram bastante do observado nas pesquisas. Na taxa de aprovação calculada, os dois candidatos a aumentaram constantemente ao longo dos meses, porém nas pesquisas realizadas Ciro Gomes decresceu na intenção de votos e Felipe d'Avila permaneceu constante.

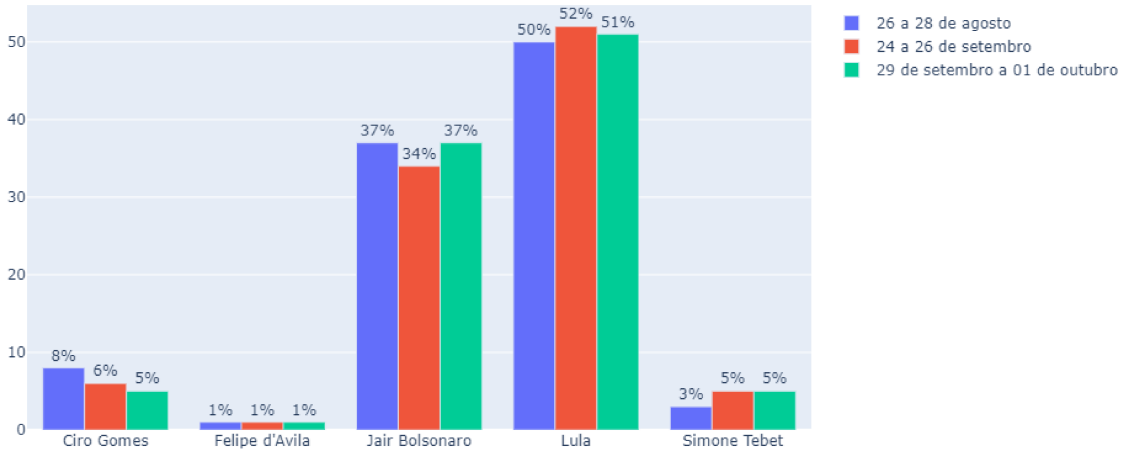

Figura 34 – Votos válidos estimados para presidente

5 Conclusão

Este trabalho teve como objetivo verificar se o desempenho de candidatos à presidência da República nas eleições brasileiras de 2022 está relacionado com a sua popularidade nas redes sociais. Para isso, o *Twitter* foi escolhido como fonte de coleta dos dados, devido à sua facilidade e popularidade. Além disso, foram utilizadas as seguintes técnicas para extração de conhecimento textuais a fim de identificar o sentimento presente em cada *tweet*: coleta de dados, criação da base de dados rotulada, pré-processamento de textos, construção de nuvens de palavras, classificação dos textos através do algoritmo SVM e validação dos resultados. Ao fim, o sentimento dos *tweets* foi contrastado com o resultado das eleições.

Ao final da coleta de dados, 6.778.481 *tweets* foram extraídos da rede social, sendo que o mês de Agosto se destacou por obter o maior número de *tweets* coletados, devido ao lançamento oficial das candidaturas de cada partido. Apesar de os dados terem sido coletados com base na região disponibilizada pelo usuário em seu perfil, essa função presente na API de coleta ainda é muito imprecisa e instável, de acordo com sua documentação, por ter sido uma função desabilitada oficialmente pela plataforma.

Um dos objetivos secundários deste estudo foi analisar a viabilidade do *transfer learning* para problemas de classificação relacionados à eleições presidenciais. Observou-se que os resultados obtidos com o método de *transfer learning* foram um pouco abaixo dos obtidos com o classificador treinado com a base de dados classificada manualmente pela autora. Por isso, optou-se por não utilizá-lo, apesar de apresentar resultados semelhantes.

Em relação ao classificador, o resultado obtido poderia ter sido melhor se houvesse um avanço no pré-processamento, como identificação de *tweets* realizados por *bots* , aprimoramento da lista de *stopwords* e correção de palavras/gírias. Além disso, outra técnica que poderia ter sido implementada é a identificação e remoção de textos jornalísticos que não apresentam opiniões sobre candidatos. Métodos mais robustos para classificação de textos como o BERT (*Bidirectional Encoder Representations from Transformers*) poderiam trazer melhores resultados.

Outro ponto interessante a ser comentado em relação à classificação, é em relação aos comentários coletados. Foi identificado que em um mesmo *tweet*, o usuário pode ter sentimentos conflitantes para candidatos diferentes, como por exemplo, falar mal de um candidato para enaltecer outro que também é mencionado. Nesse caso, como o comentário possui palavras de cunho negativo, o classificador entende que o sentimento expresso é negativo. Outro caso bastante comum foi, elogiar um candidato e no mesmo comentário falar mal de outro, encontrando portanto palavras em tons negativos e positivos em um

mesmo *tweet*. Visando esse problema, novas pesquisas mais avançadas têm surgido com o propósito de analisar o sentimento expresso no comentário mais detalhadamente. Segundo Kauer (2016), a Análise de Sentimentos Baseada em Aspectos busca obter o máximo de detalhes sobre a entidade que está sendo classificada, separando para cada comentário o alvo da opinião, sentimento atribuído e categoria correspondente. Dessa forma, é possível realizar a separação de opiniões sobre candidatos em um mesmo comentário. Assim, técnicas mais avançadas como essa poderiam trazer melhores resultados.

A partir da classificação dos *tweets* coletados, foi feita uma comparação com as pesquisas eleitorais realizadas nos meses de agosto, setembro e outubro, assim como o resultado oficial do primeiro turno. Alguns paralelos puderam ser traçados principalmente em relação às pesquisas eleitorais, como o aumento da popularidade da candidata Simone Tebet nas redes sociais e no índice de intenção de votos e o aumento da popularidade do candidato Jair Bolsonaro ao longo dos meses.

Apesar disso, não foi encontrada uma associação significativa entre a popularidade de um candidato e seu desempenho na eleição. Uma possível explicação para isso, algo notado e comentado anteriormente, é a grande diferença de dados coletados entre os candidatos. O candidato Felipe d'Avila, por exemplo, apesar de ter sido o mais bem avaliado, obteve 9.772 *tweets* coletados, o que representa apenas 0,001% da base de dados coletada. Isso pode significar que apesar de ser bem avaliado pelo seu público, Felipe d'Avila não é bem conhecido pela maioria dos brasileiros, impactando diretamente seu desempenho na eleição.

Por fim, observa-se também uma divergência entre a quantidade de *tweets* coletados em determinadas regiões e o tamanho do seu colégio eleitoral. Um exemplo seria a região Centro-Oeste, que no caso do estudo foi a 2° maior região em termos de dados coletados, porém, apenas 7% dos eleitores brasileiros votam na região. Outro caso observado foi em relação à região Sul. Representando 14% dos eleitores brasileiros, sendo o 3° maior colégio eleitoral do país, foi a região com menos dados coletados, representando 4% dos *tweets* totais.

5.1 Contribuições

As principais contribuições deste estudo foram:

1. Criação de uma base de dados composta por 6.778.481 de *tweets* a respeito da eleição presidencial de 2022. A base contém o *tweet* coletado, data e região do usuário.

2. Construção de uma base de dados sobre os candidatos que disputaram a eleição presidencial de 2022, composta por 1000 tweets rotulados manualmente com os

sentimentos positivo, negativo ou neutro. Essa base pode ser usada como conjunto de treinamento para algoritmos de aprendizagem supervisionada.

3. Indicação do algoritmo *SVM* como promissor para técnicas de análise de sentimentos e tarefas de classificação de dados

4. Criação de um projeto de análise de sentimentos, disponível para ser utilizado em trabalhos futuros[1].

5. Desenvolvimento de um estudo experimental que contrasta o resultado das eleições para presidente do ano de 2022 e a popularidade dos candidatos no *Twitter*.

5.2 Trabalhos futuros

Neste trabalho, foram aplicadas diversas técnicas de pré-processamento, porém existem outras que poderiam ser utilizadas também, como a identificação de *bots* e remoção de textos jornalísticos. Além disso, as *hashtags* presentes nos *tweets* poderiam ser removidas também a fim de avaliar o impacto nos indicadores dos classificadores.

Para uma classificação mais minunciosa, os *tweets* de cada candidato poderiam ser desconsiderados se comentassem outros candidatos ou separados em categorias diferentes. Foi observado no estudo, que apesar de conter o nome de um candidato em questão, existem *tweets* que fazem uma comparação dele com outro candidato, o elogiando/reprovando. Dessa forma, o texto está contido na base de um candidato, sendo que é referente ao sentimento que o usuário possui de outro candidato, gerando portanto confusão nas taxas de aprovação/rejeição.

Além disso, como citado anteriormente, a quantidade de *tweets* coletados influencia no desempenho que o candidato possui na eleição. Em trabalhos futuros, seria interessante atribuir um peso para os candidatos com base na quantidade de *tweets* colhidos, dessa forma, candidatos que possuem uma base pequena não se destacariam tanto quanto outros mais comentados. Novos algoritmos e métodos de classificação de sentimentos, como a Análise de Sentimentos Baseada em Contexto, deveriam ser investigados, a fim de melhorar o desempenho desta tarefa.

Por fim, seria interessante também analisar o sentimento dos usuários da rede social após o resultado final do 2° turno, a fim de analisar a popularidade do presidente eleito.

[1] Projeto completo disponível em bit.ly/projetoAnaliseSentimentos

Referências

AHLGREN, M. **50 + TWITTER ESTATÍSTICAS E FATOS PARA 2022**. 2022. Urlhttps://www.websiterating.com/pt/research/. Citado na página 14.

ALMEIDA, R. d. Q. Fake news: arma potente na batalha de narrativas das eleições 2018. **Ciência e Cultura**, Sociedade Brasileira para o Progresso da Ciência, v. 70, n. 2, p. 9–12, 2018. Citado na página 15.

ALVES, A. L. F.; BAPTISTA, C. D. S.; FIRMINO, A. A.; OLIVEIRA, M. G. d.; PAIVA, A. C. d. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup. In: **Proceedings of the 20th Brazilian Symposium on Multimedia and the Web**. [S.l.: s.n.], 2014. p. 123–130. Citado na página 21.

ALVES, F. de F. **Marketing político e eleitoral : um estudo sobre as estratégias e ferramentas necessárias para a construção de uma campanha política**. 48 p. — UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, Rio de Janeiro, RJ, Brasil, 2018. Citado na página 15.

ANACLETO, A. C. d. S. **Aplicação de Técnicas de Data Mining em Extracção de Elementos de Documentos Comerciais**. Dissertação (Mestrado) — Faculdade de Economia da Universidade do Porto, 2010. Citado na página 21.

ARANHA, C. N. Processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. **Rio de Janeiro: PUC-Rio**, 2007. Citado na página 18.

ARAÚJO, M.; GONÇALVES, P.; BENEVENUTO, F.; CHA, M. Métodos para análise de sentimentos no twitter. In: SN. **Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)**. [S.l.], 2013. p. 19. Citado na página 9.

ATTUX, R. C. **Predição dos resultados das eleições 2014 para presidente do Brasil usando dados do Twitter**. 47 p. — Universidade Federal de Uberlândia, Uberlândia, MG, Brasil, 2017. Citado 2 vezes nas páginas 23 e 25.

BENEVENUTO, F.; ALMEIDA, J.; SILVA, A. Coleta e análise de grandes bases de dados de redes sociais online. In: **Jornadas de Atualização em Informática (JAI)**. [S.l.: s.n.], 2012. Citado 2 vezes nas páginas 14 e 19.

BERRAR, D. Cross-validation. In: ELSEVIER (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. [S.l.: s.n.], 2018. p. 542–545. Citado na página 21.

BORGES, A. Nacionalização partidária e estratégias eleitorais no presidencialismo de coalizão. **Dados**, SciELO Brasil, v. 58, p. 651–688, 2015. Citado na página 13.

CAJADO, A. F. R.; DORNELLES, T.; PEREIRA, A. C. **ELEIÇÕES no Brasil: uma história de 500 anos**. [S.l.]: Tribunal Superior Eleitoral, 2014. 99 p. Citado na página 12.

CARVALHO, L. B. D. A democracia frustrada: fake news, política e liberdade de expressão nas redes sociais. **internet&sociedade**, v. 1, n. 1, p. 172–199, 2020. Citado na página 9.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T. P.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0: Step-by-step Data Mining Guide**. SPSS, 2000. Disponível em: <https://books.google.com.br/books?id=po7FtgAACAAJ>. Citado 2 vezes nas páginas 17 e 19.

CRISTIANI, A.; LIEIRA, D.; CAMARGO, H. A sentiment analysis of brazilian elections tweets. In: **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2020. p. 153–160. ISSN 2763-8944. Disponível em: <https://sol.sbc.org.br/index.php/kdmile/article/view/11971>. Citado 8 vezes nas páginas 6, 23, 25, 28, 29, 31, 43 e 45.

DOURADO, T. M. S. G. **Fake News na eleição presidencial de 2018 no Brasil**. Tese (Doutorado) — Instituto de Humanidades, Artes e Ciências Professor Milton Santos, 2020. Citado na página 15.

DUTRA, D. A. M.; FRANCISCO, E. de R. Text mining: Análise de sentimentos nas eleiÇões 2018. In: **Congresso Transformação Digital 2018**. [S.l.: s.n.], 2018. Citado 2 vezes nas páginas 23 e 25.

FARIA, V. S. D. Eleições no império: considerações sobre representação política no segundo reinado. **Simpósio Nacional de História**, p. 1–17, 2013. Citado na página 12.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>. Citado na página 16.

FELONIUK, W. O desenvolvimento normativo do direito eleitoral no período colonial brasileiro. In: **História do direito**. [S.l.: s.n.], 2014. p. 46–72. ISBN 978-85-68147-34-4. Citado na página 12.

FIGUEIREDO, E. B.; CATINI, R. de C.; MENDES, L. M. Mineração de textos: Análise de sentimento em redes sociais-revisão sistemática. **Anais do WCF**, v. 5, p. 24–29, 2018. Citado na página 18.

FILHO, R. M.; ALMEIDA, J. M.; PAPPA, G. L. Pesquisa eleitoral em redes sociais: Inclusao da análise de novas dimensoes. In: SBC. **Anais do III Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2014. p. 164–175. Citado na página 15.

FLORES, P. **Redes Sociais e TV: qual o peso de cada meio nas eleições de 2018**. 2018. Urlhttps://www.nexojornal.com.br/expresso/2018/03/18/Redes-sociais-e-TV-qual-o-peso-de-cada-meio-nas-elei%C3%A7%C3%B5es-de-2018. Citado na página 15.

FONSECA, B.; SANTINO, M. **Na manhã do 7 de setembro, críticas a Bolsonaro dominam trending topics no Twitter**. 2022.

Urlhttps://apublica.org/sentinela/2022/09/na-manha-do-7-de-setembro-criticas-a-bolsonaro-dominam-trending-topics-no-twitter/. Citado 2 vezes nas páginas 9 e 15.

FUSHIKI, T. Estimation of prediction error by using k-fold cross-validation. **Statistics and Computing**, Springer, v. 21, n. 2, p. 137–146, 2011. Citado na página 21.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining**. [S.l.]: Gulf Professional Publishing, 2005. Citado na página 22.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. **arXiv preprint arXiv:2008.05756**, 2020. Citado na página 22.

HAN, J.; PEI, J.; TONG, H. **Data mining: concepts and techniques**. [S.l.]: Morgan kaufmann, 2022. Citado 2 vezes nas páginas 16 e 20.

INSPER. **MUNDO SE APROXIMA DA MARCA DE 5 BILHÕES DE USUÁRIOS DE INTERNET, 63% DA POPULAÇÃO**. 2022. Urlhttps://www.insper.edu.br/noticias/mundo-se-aproxima-da-marca-de-5-bilhoes-de-usuarios-de-internet-63-da-populacao/. Citado na página 14.

JARDELINO, F.; CAVALCANTI, D. B.; TONIOLO, B. P. A proliferação das fake news nas eleições brasileiras de 2018. **Comunicação Pública**, v. 15, n. 28, Set. 2021. Disponível em: <https://journals.ipl.pt/cpublica/article/view/99>. Citado na página 15.

JUNIOR, G. d. B. V.; LIMA, B. N.; PEREIRA, A. de A.; RODRIGUES, M. F.; OLIVEIRA, J. R. L. de; SILIO, L. F.; CARVALHO, A. dos S.; FERREIRA, H. R.; PASSOS, R. P. Determinação das métricas usuais a partir da matriz de confusão de classificadores multiclasses em algoritmos inteligentes nas ciências do movimento humano. **Revista CPAQV-Centro de Pesquisas Avançadas em Qualidade de Vida-CPAQV Journal**, v. 14, n. 2, 2022. Citado na página 22.

JUNIOR, J. R. C. Desenvolvimento de uma metodologia para mineração de textos. **Pontificia Universidad Catolica de Rio de Janeiro: Rio de janeiro, Brasil**, 2007. Citado 3 vezes nas páginas 19, 20 e 22.

KAUER, A. U. **Análise de sentimentos baseada em aspectos e atribuições de polaridade**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2016. Citado na página 59.

LEMOS, E. M. d. S. **Marketing político e mídias digitais: um estudo sobre a influência das redes sociais na política brasileira**. 2019. Disponível em: <https://bdm.unb.br/handle/10483/25104>. Citado na página 15.

LÓPEZ, C. **DATA MINING. The CRISP-DM METHODOLOGY. The CLEM language and IBM SPSS MODELER**. Lulu.com, 2021. ISBN 9781008981652. Disponível em: <https://books.google.com.br/books?id=pYAnEAAAQBAJ>. Citado na página 16.

MARQUES, F. P. J. A.; SAMPAIO, R. C. Internet e eleições 2010 no brasil: rupturas e continuidades nos padrões mediáticos das campanhas políticas online. **Galáxia. Revista do Programa de Pós-Graduação em Comunicação e Semiótica**, n. 22, 2011. Citado na página 9.

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. Tese (Doutorado) — Universidade de São Paulo, 2003. Citado na página 19.

MARTINS, C. A.; MONARD, M. C.; MATSUBARA, E. T. Uma ferramenta computacional para auxiliar no pré-processamento de textos. In: **Anais do XXIII Congresso da Sociedade Brasileira de Computaçao-IV Encontro Nacional de Inteligência Artificial (ENIA), Campinas, SP**. [S.l.: s.n.], 2003. v. 6. Citado 2 vezes nas páginas 19 e 20.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams engineering journal**, Elsevier, v. 5, n. 4, p. 1093–1113, 2014. Citado na página 18.

MELO, C. R. F. D. Eleições presidenciais, jogo aninhados e sistema partidário no brasil. **Revista Brasileira de Ciência Política**, Universidade de Brasília, Instituto de Ciência Política, v. 4, p. 13, 2010. Citado na página 12.

MIGUEL, L. F.; TOKARSKI, F. M. B.; MOTA, F. F. Mídia, eleições e pesquisa de opinião no brasil (1989-2010): um mapeamento da presença das pesquisas na cobertura eleitoral. **Revista ComPolítica**, Associação Brasileira de Pesquisadores em Comunicação e Política, 2011. Citado na página 13.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007. Citado na página 18.

MORI, L. **Por que ex-aliados do presidente adotaram termo 'bolsopetismo' para atacar governistas**. 2020. Urlhttps://www.bbc.com/portuguese/brasil-53187626. Citado na página 41.

NASCIMENTO, P.; OSIEK, B.; XEXÉO, G. Análise de sentimento de tweets com foco em notícias. **Revista Eletrônica de Sistemas de Informação**, v. 14, n. 2, 2015. Citado na página 14.

PASSOS, C. A. e E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, 2006. ISSN 1677-3071. Disponível em: <http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>. Citado na página 18.

PEREIRA, J. G. **Análise de sentimentos da populaçao brasileira em relaçao a eleiçao presidencial de 2018 através da rede social twitter**. Dissertação (B.S. thesis) — Universidade Federal do Rio Grande do Norte, 2019. Citado na página 9.

QUEIROZ, G.; ALMEIDA, L. Uma metodologia de análise de sentimentos dos candidatos as eleições presidenciais de 2018 no twitter. **Revista de Engenharia e Pesquisa Aplicada**, v. 5, p. 21–30, 04 2020. Citado 2 vezes nas páginas 24 e 25.

RECUERO, R.; ZAGO, G. Em busca das "redes que importam": redes sociais e capital social no twitter. **Líbero**, n. 24, p. 81–94, 2016. Citado na página 14.

REIS, F. W. Eleição de 2014:'país dividido'e questão social. **Em Debate**, v. 6, p. 8–1, 2014. Citado na página 13.

RESENDE, R. **Sudeste concentra 42% do eleitorado nacional**. 2022. Urlhttps://www12.senado.leg.br/radio/1/noticia/2022/09/23/sudeste-concentra-42-do-eleitorado-nacional. Citado na página 36.

RISH, I. et al. An empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46. Citado na página 20.

ROSSETTO, G. P. N.; CARREIRO, R.; ALMADA, M. P. Twitter e comunicação política: limites e possibilidades. **Compolítica**, v. 3, n. 2, p. 189–216, 2013. Citado na página 9.

ROSSINI, P. G. da C.; BAPTISTA, É. A.; OLIVEIRA, V. V. de; SAMPAIO, R. C. O uso do facebook nas eleições presidenciais brasileiras de 2014: a influência das pesquisas eleitorais nas estratégias das campanhas digitais. **Fronteiras-estudos midiáticos**, v. 18, n. 2, p. 145–157, 2016. Citado na página 13.

SALVADORI, I. L. et al. **Desenvolvimento de web apis restful semânticas baseadas em json**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2015. Citado na página 19.

SCHNEIDER, C. F. **Machine learning aplicado na previsão de resultados de partidas de futebol: um estudo de caso para comparação de diferentes classificadores**. 2018. Disponível em: <https://lume.ufrgs.br/handle/10183/179461>. Citado na página 22.

SILVA, M. P. d.; SANTOS, N. d. **A influência das redes sociais na campanha eleitoral**. 2020. 5 p. Disponível em: <https://www.unifan.edu.br/unifan/aparecida/wp-content/uploads/sites/2/2020/02/A-INFLUÊNCIA-DAS-REDES-SOCIAIS.pdf>. Citado na página 15.

SOARES, D. A. **Recurso de geolocalização do Twitter será desativado por falta de uso**. 2019. Urlhttps://www.tecmundo.com.br/redes-sociais/142947-recurso-geolocalizacao-twitter-desativado-falta-uso.htm. Citado na página 27.

SOUZA, I. **Pesquisas eleitorais: como são feitas?** 2022. Urlhttps://www.politize.com.br/pesquisas-eleitorais-como-sao-feitas/. Citado na página 13.

TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. **RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação**, v. 8, p. 53–65, 12 2011. Citado na página 14.

TSE. **Estatísicas de eleição**. 2022. Urlhttps://sig.tse.jus.br/ords/dwapr/seai/r/sig-eleicao-resultados/resultado-consolidado?p0$_a brangencia = Brasil clear = RP session =$ 12289352914359.$Citado na página 13$.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. **Journal of Big data**, SpringerOpen, v. 3, n. 1, p. 1–40, 2016. Citado na página 28.

WILLETT, P. The porter stemming algorithm: then and now. **Program**, Emerald Group Publishing Limited, 2006. Citado na página 19.

ZENHA, L. Redes sociais online: o que são as redes sociais e como se organizam? **Caderno de Educação**, v. 49, 2018. Citado na página 14.