Mineração de Opinião na rede social *X* sobre as eleições presidenciais de 2022

Bruno V. Vasconcelos[1], **Cristiane N. Targa**[1], **Carlos A. Silva**[1]

[1]Departamento de Informática – Instituto Federal de Minas Gerais (IFMG)
CEP – 34590-390 – Sabará – MG – Brasil

brunovictorvasconcelos@gmail.com,

{cristiane.targa,carlos.silva}@ifmg.edu.br

Abstract. *Social networks, commonly present in people's daily lives, have been widely used to promote products, services, and express opinions, becoming a relevant tool for searching and analyzing information on various subjects. In the political scenario, the use of social networks has intensified and has been employed as one of the primary means of communication with voters, in addition to representing a significant indicator of popularity in politics. The objective of this work is to collect X data related to the two candidates who reached the second round, Bolsonaro and Lula, and classify their posts as positive, negative, or neutral. To accomplish this, we created a database, pre-processed the data, and utilized the Naive Bayes method for post classification.*

Resumo. *As redes sociais, comumente presentes no cotidiano das pessoas, tem sido bastante utilizadas para divulgar produtos, serviços e expressar opiniões, configurando-se como uma relevante ferramenta para busca e análise de informações sobre os mais variados assuntos. No cenário político, o uso das redes sociais tem se intensificado e sido utilizado como uma das principais formas de comunicação com os eleitores, além de representar um significativo indicador da popularidade na política. O objetivo desse estudo é classificar os sentimentos dos posts da rede social X relacionado com os dois candidatos que foram para o segundo turno das eleições presidenciais brasileira de 2022, Bolsonaro e Lula, em positivos, negativos ou neutros, a fim de compreender e avaliar a percepção pública em relação a esses políticos. Para isso, construímos uma base de dados, realizamos um pré-processamento dos dados e utilizamos o algoritmo de Naive Bayes para classificar os posts.*

1. Introdução

Nas eleições de 2022, o cenário político brasileiro foi marcado por uma atividade intensa, com eleitores buscando informações sobre os candidatos em várias fontes informativas. Esse interesse público elevado se manifestou não apenas em campanhas e debates, mas também nas redes sociais, onde as discussões políticas ganharam destaque.

As redes sociais já faziam parte da rotina de muitas pessoas e, durante o período eleitoral de 2022 no Brasil, tornaram-se espaços ainda mais importantes para interação e comunicação. Estas plataformas são amplamente utilizadas e também servem como recursos para diversos trabalhos sobre mineração de opinião. Redes sociais como *Facebook*, *Instagram* e o antigo *Twitter*, que a partir de julho de 2023 passou a se chamar

X, fornecem ambientes onde vários tipos de conteúdo podem ser criados e compartilhados. Essas redes sociais têm se tornado cada vez mais populares, oferecendo aos usuários diversas maneiras de expressar suas opiniões. Na rede social *X*, por exemplo, encontramos um ambiente onde distintos assuntos ganham relevância no cenário nacional e mundial, o que gera uma enorme quantidade de dados para coleta e análise. De acordo com [Bragança and Braga 2023], no Brasil existem aproximadamente 171,5 milhões de usuários ativos em redes sociais, o que representa 79,9% da população brasileira. Esse número reflete um crescimento de 14,3% ou 21 milhões de usuários entre 2021 e 2022[1].

O aumento do número de usuários em plataformas de redes sociais tem impulsionado uma elevação na geração e utilização de dados em escala global nos últimos anos. De acordo com [Statista 2022], em 2025 este quantitativo deve ultrapassar a faixa de 180 zettabytes. Neste cenário de abundância de dados, as informações textuais têm sido amplamente utilizadas para diversos propósitos, tais como gestão de serviços [Kumar et al. 2021], prevenção de crimes cibernéticos [Andleeb et al. 2019], e caracterização de clientes por meio de redes sociais [He et al. 2019]. Segundo Liu [Liu 2010], as informações textuais podem ser classificadas em dois tipos: fatos e opiniões. Enquanto os fatos denotam expressões objetivas, as opiniões referem-se a expressões subjetivas que descrevem sentimentos ou avaliações a respeito de entidades ou eventos. Por exemplo, um fato relacionado ao contexto político seria "*O presidente foi eleito com mais de 50% dos votos nas eleições*", enquanto uma opinião seria "*O presidente está fazendo um ótimo trabalho na administração do país*". Neste trabalho, a diferença entre fatos e opiniões tem um peso muito significativo, pois analisamos apenas os cenários contendo opiniões.

Neste estudo, abordou-se a análise de *posts* relacionados às eleições presidenciais de 2022, empregando critérios específicos para a seleção dessas publicações. Foram consideradas apenas as publicações que incluíam as palavras "Lula" ou "Bolsonaro" e que estavam dentro das datas limite correspondentes ao primeiro e segundo turno das eleições. A busca foi limitada aos 5 mil primeiros *posts* do dia para um controle mais eficiente da base de dados, seguida por uma limpeza dos dados para facilitar a classificação. Utilizando o algoritmo de *Naive Bayes*, a análise revelou que a grande maioria dos *posts* expressa opiniões negativas sobre os candidatos. Esse tipo de análise, conhecida como mineração de texto ou análise de opinião, visa quantificar o valor emocional contido em palavras ou textos, proporcionando uma compreensão mais profunda das opiniões e emoções expressas [Redhu et al. 2018]. Quando essas informações são obtidas a partir de redes sociais e devidamente analisadas, podem contribuir para a compreensão, explicação e até mesmo a predição de complexos fenômenos sociais [Benevenuto et al. 2015].

O artigo está organizado da seguinte forma. Na seção 1 é realizada uma introdução ao problema de pesquisa, destacando sua relevância e apresentando a proposta de solução adotada. A seção 2 aborda trabalhos relacionados e traz uma revisão bibliográfica sobre o tema. Nas seções 3 e 4 são detalhadas a metodologia empregada e a descrição minuciosa dos passos do desenvolvimento. A seção 5 discute os resultados alcançados com base na metodologia aplicada. Por fim, na seção 6, apresenta-se a conclusão do trabalho, sintetizando os principais pontos abordados e possíveis direções futuras de pesquisa.

[1]https://www.insper.edu.br/noticias/mundo-se-aproxima-da-marca-de-5-bilhoes-de-usuarios-de-internet-63-da-populacao/

2. Trabalhos relacionados

No contexto da análise de sentimentos presentes em redes sociais, ferramentas de classificação automática têm sido amplamente utilizadas. Entre elas, destacam-se o algoritmo de *k-means*, mencionado por [Queiroz and Almeida 2020], e o algoritmo de *Naive Bayes* utilizado por [Dutra and Francisco 2018], [Pereira 2019] e [Souza et al. 2022]. Esses algoritmos têm a capacidade de examinar um texto e automaticamente categorizar o sentimento do autor em relação a um determinado tópico.

Algumas ferramentas para análise de sentimentos realizam uma classificação binária, rotulando o texto como positivo ou negativo. Por exemplo, [Massetti and Fernandes 2017] aplicou o algoritmo de *Naive Bayes* para analisar *posts* direcionados aos candidatos à prefeitura de São Luís em 2016. Outras ferramentas adotam uma abordagem ternária, classificando o texto como positivo, negativo ou neutro, como evidenciado por [Chandak 2022] em sua análise das eleições presidenciais de 2020 nos EUA, utilizando o algoritmo *Bag of Words*.

O estudo de [Massetti and Fernandes 2017] teve como objetivo classificar as opiniões dos eleitores sobre um debate entre os candidatos à prefeitura de São Luís utilizando *Naive Bayes*, extraídas do *Twitter*[2] em 2016. Em suas análises, foi empregada a classificação binária, destacando resultados exclusivamente positivos ou negativos. Os resultados revelaram a insatisfação dos eleitores usuários do *Twitter* em relação ao debate.

Já o trabalho de [Caetano et al. 2017] propõe identificar e analisar a homofilia política entre os usuários do *Twitter* ultilizando o algoritmo de *SentiStrength* ao longo da campanha presidencial americana de 2016. A homofilia dá significado a indivíduos que apresentam características parecidas, quando se tem a mesma origem, Para isso, categorizam suas análises finais como positivas, negativas, neutras e não manifestadas. Seus resultados mostraram que o maior grau de homofilia ocorreu entre usuários que publicaram *tweets* com sentimentos negativos em relação a Donald Trump. No entanto, devido à complexidade de entender, interpretar e classificar algumas declarações, outros autores preferem uma abordagem mais binária, focando apenas em resultados positivos ou negativos. Essa classificação binária proporciona um resultado mais direto, como evidenciado no estudo de [Queiroz and Almeida 2020] que tinha como objetivo apresentar uma metodologia para analisar os sentimentos expressos no *Twitter* em relação aos candidatos nas eleições presidenciais de 2018 no Brasil. Os resultados desse estudo apontaram um grau de similaridade entre os candidatos avaliados.

Nas eleições presidenciais de 2020 dos Estados Unidos, [Chandak 2022] também optou por classificar de forma não binária seus dados. De acordo com os autores, a análise de sentimentos representa um desafio significativo. Eles concluíram que, após o processamento dos dados, os *tweets* de Trump se tornaram mais positivos, enquanto os *tweets* de Biden se tornaram mais negativos. Isso ocorre porque os usuários de mídias sociais tendem a ser bastante expansivos e livres em suas postagens, o que pode dificultar a interpretação por computadores. Portanto, uma etapa de pré-processamento é crucial para preparar o texto para que possa ser decifrado por máquinas.

[Matos et al. 2020] identificou uma quantidade significativa de sentimentos negativos ao utilizar o algoritmo *Orange Canvas* nos *tweets* recuperados durante o período

[2]Em julho de 2023 a rede social *Twiter* passou a ser chamada de *X*.

eleitoral das eleições presidenciais do Brasil de 2018. Suas pesquisas também concluíram que o *Twitter* é um espaço onde os usuários frequentemente expressam seus sentimentos em relação às eleições. No entanto, apesar dessas observações, o estudo não conseguiu antecipar o resultado final das eleições. Acredita-se que isso ocorreu devido à insatisfação com os candidatos e à crescente homofilia política nos últimos anos. A conclusão do estudo apontou que em grupos com maior homofilia, os usuários tendiam a expressar sentimentos negativos e críticas. Por outro lado, em grupos com maior heterofilia, os usuários mostraram uma tendência menor a manifestar seus sentimentos.

Para validar os resultados gerados pelos algoritmos utilizados na classificação pode-se utilizar métricas que validam a confiabilidade. Em [Massetti and Fernandes 2017] as métricas de precisão, acurácia e *recall* foram utilizadas. Quando analisamos o modelo de teste, podemos utilizar métricas que ajudam a ter uma maior confiabilidade no algoritmo. As métricas de precisão, acurácia e *recall* são importantes para avaliar a qualidade de um modelo de aprendizado de máquina. A precisão mede a proporção de instâncias positivas corretamente identificadas. A acurácia representa a proporção de previsões corretas em relação ao total de instâncias avaliadas, refletindo a precisão geral do modelo, e por fim, o *recall* avalia a capacidade do algoritimo de identificar instâncias positivas, sendo a proporção de instâncias positivas corretamente identificadas em relação ao total de instâncias reais positivas.

3. Metodologia

As etapas de desenvolvimento deste trabalho são ilustradas na Figura 1.

Figura 1. Etapas do desenvolvimento.
Fonte: Autores.

Inicialmente, foram coletados os *posts* da rede social *X*. Em seguida, os dados coletados foram pré-processados, o que envolveu a remoção de informações desnecessárias e a limpeza dos dados brutos. Os dados foram categorizados utilizando o algoritmo *Naive Bayes*, que permitiu identificar o sentimento predominante em cada *post*, sendo positivos, negativos ou neutros.

Usamos a biblioteca *snscrape*[3] para realizar a coleta dos *posts* aplicando filtros de busca específicos. Foram realizadas duas coletas, uma durante o primeiro turno e outra

[3]https://github.com/JustAnotherArchivist/snscrape

durante o segundo turno das eleições presidenciais de 2022, focando em Bolsonaro e Lula, o que resultou em aproximadamente 700 mil *posts*.

Após a coleta dos *posts*, foi implementado um algoritmo utilizando linguagem *Python* na plataforma *Google Colab*[4] com intuito de limpar e analisar os dados que seriam utilizados para treinamento do modelo. Esses *posts* contém informações gerais a respeito do governo de Minas Gerais em 2017.Além disso, os *posts* desta base foram uma fonte valiosa de dados, pois compartilham o mesmo contexto político, o que enriqueceu bastante a análise. Treinamos o classificador *Naive Bayes* usando *posts* já classificados como positivos, negativos ou neutros. Foi ultilizada a classificação não binária, pois nos permite uma representação mais precisa e abrangente dos sentimentos expressos em cada *post*. Além disso, em diversas situações, as emoções expressadas não se restringem rigidamente a classificações apenas "positivas" ou "negativas". Ao incorporar classificações neutras, é possível ter uma base muito mais ampla de sentimentos expressos.

No contexto deste trabalho, para reforçar a robustez dos nossos resultados, adotamos a metodologia de validação cruzada. Essa técnica é empregada para avaliar o desempenho de conjuntos de treinamento, abordando a questão da aleatoriedade na divisão dos dados em conjuntos de treinamento e teste, o que pode impactar os resultados. A validação cruzada, também usada por [Chaudhry et al. 2021], consiste na divisão da base de dados em várias partes de tamanhos iguais. Essas partes são utilizadas para treinamento em diferentes combinações, definindo o tamanho de cada conjunto de treinamento, no caso deste trabalho utilizamos dez partes, enquanto o restante é reservado para fins de teste. Esse processo é repetido até que todas as partes tenham sido empregadas para treinamento.

4. Desenvolvimento

Foi realizada a coleta dos dados para a construção da base utilizando um algoritmo desenvolvido em *Python* com a biblioteca *snscrape*[5]. Em seguida, passamos a base para o algoritmo criado no *Google Colab*[6] utilizado para realizar a limpeza e análise dos dados coletados. Depois disso, realizou-se a limpeza dos dados, removendo *links*, vírgulas, pontos e algumas outras *stopwords* que são palavras muito utilizadas, mas que não tem muito peso para nossa análise como "o", "a", "e", "de" e outras. Para o treinamento, foi utilizada uma base com 8199 *posts* já classificados relacionados ao governo do estado de Minas Gerais no ano de 2017, sendo estes *posts* obtidos do github *Stack Tecnologias*[7]. Essa coleta está focada nas opiniões expressas no antigo *Twitter* em relação ao governo do estado de Minas Gerais durante o ano de 2017. Os *tweets* disponibilizados já estão classificados, cada frase com sua respectiva polaridade, podendo elas serem "positivo", "neutro" ou "negativo". Após o treinamento, aplicamos o classificador de *Naive Bayes*[8] utilizando a biblioteca *SKLearn* para classificar os *posts* como positivos, negativos ou neutros.

Para a execução da primeira fase de desenvolvimento ilustrada na Figura 1, desenvolveu-se um programa implementado na linguagem de programação *Python 3.10* e

[4]https://colab.google/

[5]https://pypi.org/project/snscrape/

[6]https://colab.google/

[7]https://github.com/stacktecnologias/stack-repo/blob/master/Tweets_Mg.csv

[8]https://scikit-learn.org/stable/modules/naive_bayes.html

utilizou-se biblioteca chamada *snscrape*. Essa biblioteca foi usada para coletar os dados do *X*, pois suporta várias plataformas de mídia social, incluindo *X, Reddit* e *Instagram*. Com o *snscrape*, é possível personalizar os filtros de busca para uma pesquisa específica, o que torna mais fácil obter os resultados finais de nossa coleta. Podemos filtrar a data dos períodos eleitorais e os termos relacionados aos nomes de cada presidente, o que contribui significativamente para a precisão e eficiência da pesquisa. Essa biblioteca se baseia em uma API simples que permite que os usuários extraiam *posts*, comentários e postagens relevantes com base em palavras-chave, *hashtags* ou usuários específicos. Além disso, oferece recursos avançados, como a possibilidade de coletar dados em um intervalo de datas específico. Outras bibliotecas testadas, como a própria API do *X*, colocam limitações na data de coleta, não permitindo a coleta de dias anteriores ao período de 8 dias passados. Foram realizados dois ciclos de coleta de *posts* conforme a Tabela 1.

Tabela 1. Caracterização das coletas de dados.

Coleta	Intervalo	Significado
1ª	16 de agosto a 30 de setembro 2022	Primeiro Turno
2ª	03 a 28 de outubro 2022	Segundo Turno

A primeira coleta foi realizada durante a campanha do primeiro turno. Segundo o Tribunal Superior Eleitoral (TSE), a campanha eleitoral para a presidência do Brasil começou oficialmente no dia 16 de agosto de 2022 e terminou dia 30 de setembro de 2022, totalizando 45 dias corridos[9]. A segunda coleta foi realizada durante a campanha do segundo turno, entre os dias 03 a 28 de outubro de 2022. Além disso, a coleta dos *posts*, tanto na primeira quanto na segunda coleta, foi limitada aos candidatos a presidência que passaram para o segundo turno das eleições presidenciais de 2022, a saber, Bolsonaro e Lula. Para cada candidato foram coletados cinco mil *posts* por dia ao longo de 45 dias, que resultaram em aproximadamente 700 mil *posts*, um total aproximadamente de 2 GB em arquivos. Esta limitação diária não apenas se alinha à necessidade de controlar o tamanho da base de dados, mas também considera as limitações do *hardware* utilizado na coleta. Se não houvesse tal limitação diária, o tempo de coleta de dados da base poderia se estender indefinidamente. Ao impor essa restrição diária, garantimos que a base de dados inclua *posts* de todos os dias da campanha eleitoral, proporcionando uma representação mais abrangente do período em questão.

Para avaliar o tamanho da base de dados utilizada neste estudo em comparação com as bases citadas na literatura e usadas em trabalhos semelhantes e relevantes para o tema da pesquisa, a Tabela 2 apresenta os autores do trabalho (Bases), a quantidade correspondente de *posts* da base utilizada e a proporção da base da literatura em relação à base construída neste artigo. Considerando a cobertura das eleições de 2022, uma base de 700 mil *posts* é bastante significativa, especialmente quando comparada ao tamanho das bases utilizadas em outros estudos. Mesmo não sendo tão ampla quanto a base utilizada em [Dutra and Francisco 2018], ainda constitui uma fonte de dados considerável.

[9]https://www.tse.jus.br/comunicacao/noticias/2022/Janeiro/confira-as-principais-datas-do-calendario-eleitoral-de-2022

Tabela 2. Comparação entre bases usadas na fundamentação teórica.

Bases	***Posts***	**Proporção em relação a base usada**
[Dutra and Francisco 2018]	1.204.036	172,01%
[Queiroz and Almeida 2020]	4.608	0,66%
[Massetti and Fernandes 2017]	570	0,08%
[Matos et al. 2020]	2000	0,29%
[Suter et al. 2019]	108.893	15,56%

5. Análise dos Resultados

Com os dados tratados a partir das coletas realizadas de primeiro e segundo turno das eleições, e após o treinamento do modelo para a classificação dos *posts* coletados, fez-se a categorização dos sentimentos. A Figura 2 apresenta um gráfico comparativo entre os presidenciáveis Lula e Bolsonaro com a classificação dos *posts* no primeiro e no segundo turno.

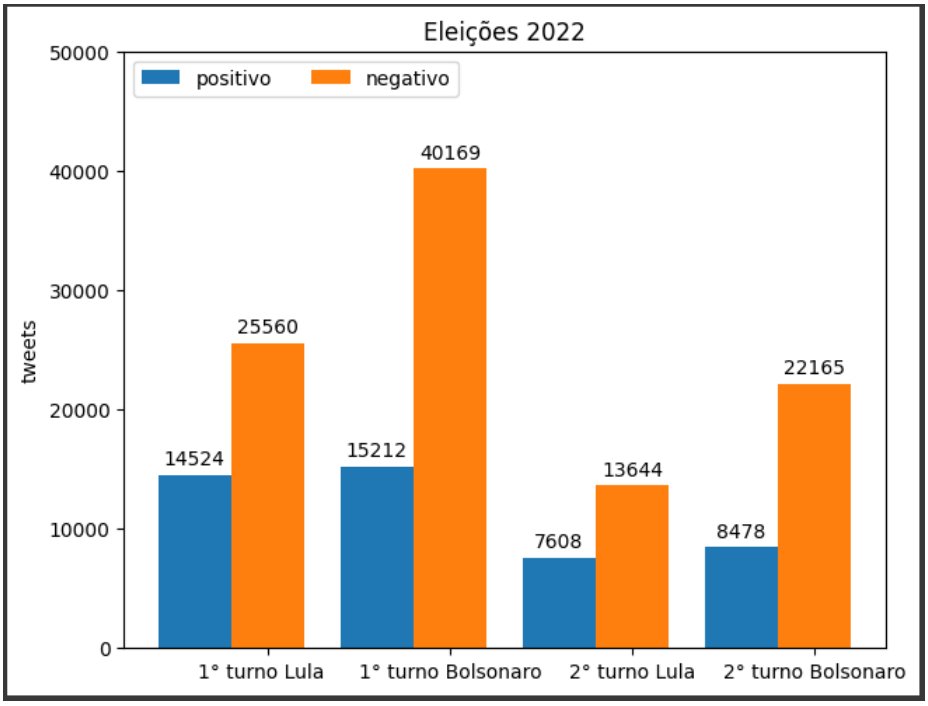

Figura 2. Comparativo de *tweets* positivos e negativos do 1º e 2º turno dos candidatos presidenciáveis.
Fonte: Autores.

Vale ressaltar que nos deparamos com uma grande quantidade de *posts* classificados como neutros, o que dificultou a interpretação durante a elaboração das Figuras 3 a 6. Diante dessa situação, decidimos ajustar nossa abordagem, usando apenas os resultados positivos e negativos. Essa mudança proporcionou uma compreensão muito mais clara e precisa dos sentimentos expressos pelos usuários.

Em um primeiro momento é possível observar que ambos os candidatos receberam

mais menções negativas do que positivas, refletindo uma tendência geral de desaprovação nos *posts* e sugerindo uma falta de apoio robusto dos eleitores para qualquer um deles. No entanto, é importante destacar que uma quantidade proporcionalmente maior de *posts* negativos foi direcionada ao candidato Bolsonaro, atingindo 162,50% em comparação com os *posts* positivos relacionados a ele. Isso sugere que, em relação aos *posts* positivos, o número de *posts* negativos associados a Bolsonaro foi significativamente maior.

Durante o primeiro turno das eleições, cerca de 17,85% dos *posts* relacionados a Bolsonaro foram classificados como negativos, em contraste com apenas 6,76% que foram positivos. No caso de Lula, aproximadamente 11,11% dos *posts* foram considerados negativos e 6,31% positivos. No segundo turno, a proporção de *posts* negativos para Bolsonaro foi de cerca de 17,05%, enquanto apenas 6,52% foram positivos. Para Lula, cerca de 5,85% dos *posts* foram classificados como negativos e 10,49% como positivos. Portanto, os resultados obtidos indicam que a rejeição desempenhou um papel significativo na dinâmica das postagens.

As Figuras de 3 a 6, apresentam de forma clara a evolução de votos positivos e negativos para os candidatos em análise. As duas linhas em cada gráfico representam o número diário de *posts* positivos e negativos durante a campanha eleitoral, destacando a variabilidade dos sentimentos ao longo do tempo. Essas linhas fornecem uma percepção visual da aprovação de Lula e Bolsonaro durante o período eleitoral, permitindo a identificação de padrões e tendências no desempenho e na aceitação do público para cada candidato.

A Figura 3 apresenta um padrão comportamental interessante para os *posts* negativos e positivos para o candidato Lula no primeiro turno.

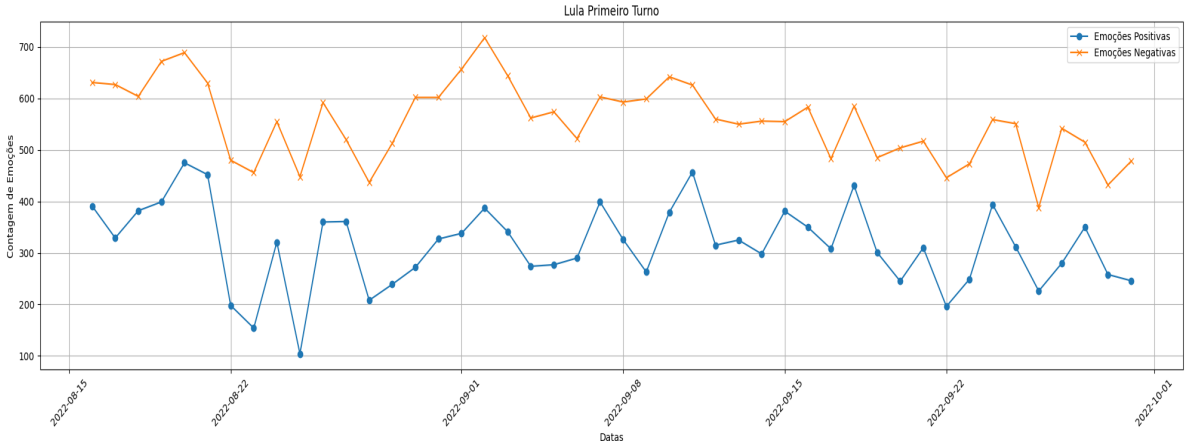

Figura 3. Primeiro turno Lula.
Fonte: Autores.

Observa-se que quando há uma queda no número de *posts* classificados como positivos, há uma tendência semelhante no número de *posts* negativos. Por exemplo, no dia 25 de agosto de 2022, após Lula conceder uma entrevista ao Jornal Nacional[10], houve uma diminuição tanto no número de *posts* positivos quanto negativos. Da mesma forma,

[10] https://www.poder360.com.br/eleicoes/leia-a-transcricao-da-entrevista-de-lula-ao-jornal-nacional/

no dia 02 de setembro de 2022, quando ocorreu um pico no número de *posts* classificados como negativos, houve também um aumento no número de *posts* positivos. Esse aumento pode ser atribuído à divulgação da pesquisa de intenção de votos após o debate entre os presidenciáveis[11], que revelou que Lula havia perdido dois pontos percentuais nas intenções de voto.

Da mesma forma que a Figura 3 (referente ao primeiro turno de Lula), a Figura 4 (relativa ao segundo turno da eleição) também demonstra um padrão de comportamento semelhante: sempre que há um crescimento no número de publicações negativas, nota-se um aumento correspondente nas publicações positivas, e o inverso também é verdadeiro.

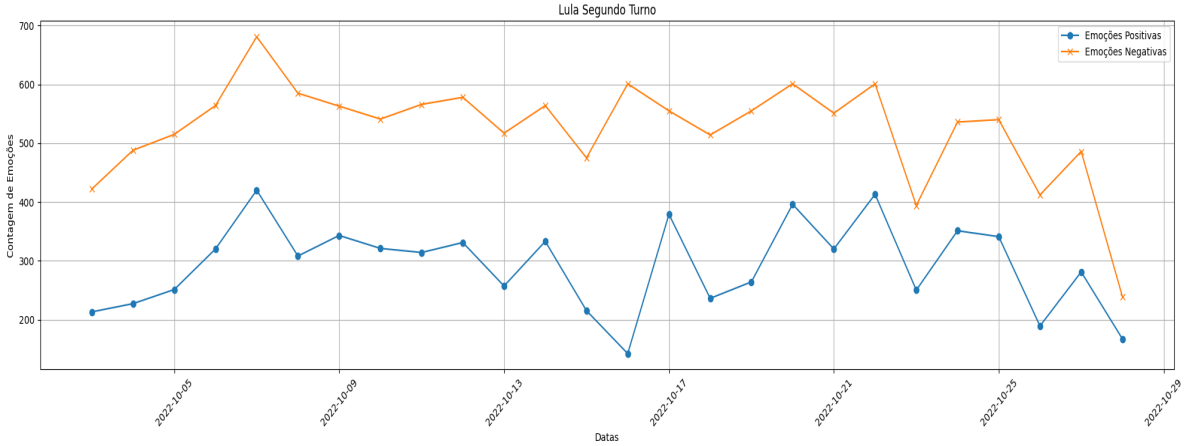

Figura 4. Segundo turno Lula.
Fonte: Autores.

No dia 07 de outubro de 2022, ocorreu um aumento no número de *posts* classificados como negativos e positivos, após Lula declarar que não revelaria nomes de ministros antes do resultado final da eleição[12]. O número de *posts* tanto negativos quanto positivos teve uma diminuição no dia 16 de outubro de 2022, data do primeiro debate do segundo turno entre Lula e Bolsonaro.

Enquanto as Figuras 3 e 4, referentes ao primeiro e segundo turno do candidato Lula, mostram o mesmo padrão de comportamento para *posts* positivos e negativos, as Figuras 5 e 6 apresentam uma dinâmica distinta para os *posts* positivos e negativos em ambos os turnos para o candidato Bolsonaro. Na Figura 5, os *posts* positivos exibem picos em alguns dias. Por exemplo, em 18 de agosto de 2022, observa-se um aumento nos *posts* negativos e uma diminuição nos positivos. Em certos dias, enquanto há um aumento no número de *posts* negativos, notamos uma quase estabilidade no número de *posts* positivos entre 25 e 30 de setembro de 2022.

[11] https://www.correiobraziliense.com.br/politica/2022/09/5033892-datafolha-apos-debate-lula-cai-para-45-ciro-e-tebet-sobem.html

[12] https://www1.folha.uol.com.br/poder/2022/10/lula-reune-aliados-de-partido-de-kassab-e-diz-que-nao-antecipara-ministros.shtml

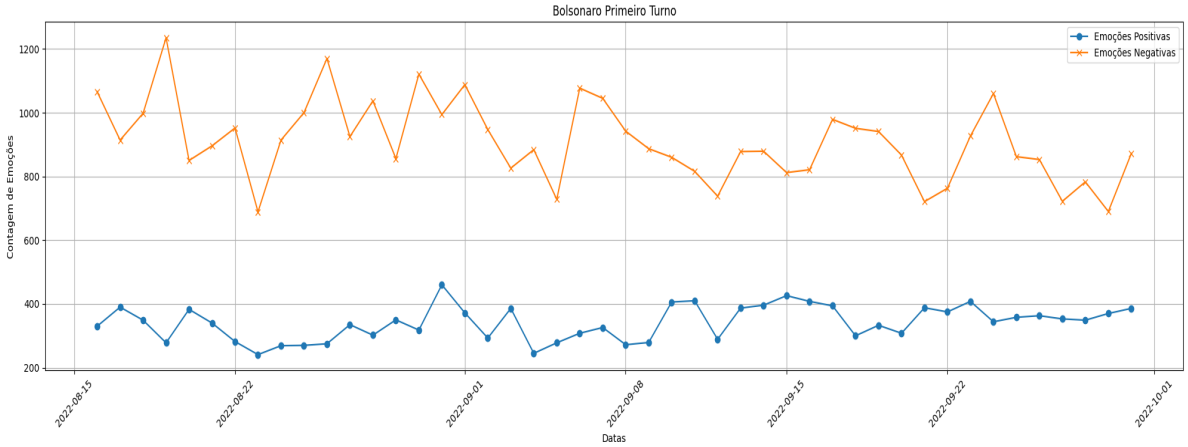

Figura 5. Primeiro turno Bolsonaro.
Fonte: Autores.

A Figura 6, que representa o segundo turno de Bolsonaro, exibe um padrão parecido, com um crescimento no número de publicações negativas e uma estabilidade nas positivas. No dia 20 de outubro de 2022, houve um grande volume de publicações categorizadas como negativas e uma queda nas positivas. Esse aumento nas publicações negativas pode ser relacionado a um *post* no perfil de Bolsonaro na plataforma X, na qual ele ironizou que Lula iria "sapecar o 22 na urna", gerando grande repercussão[13].

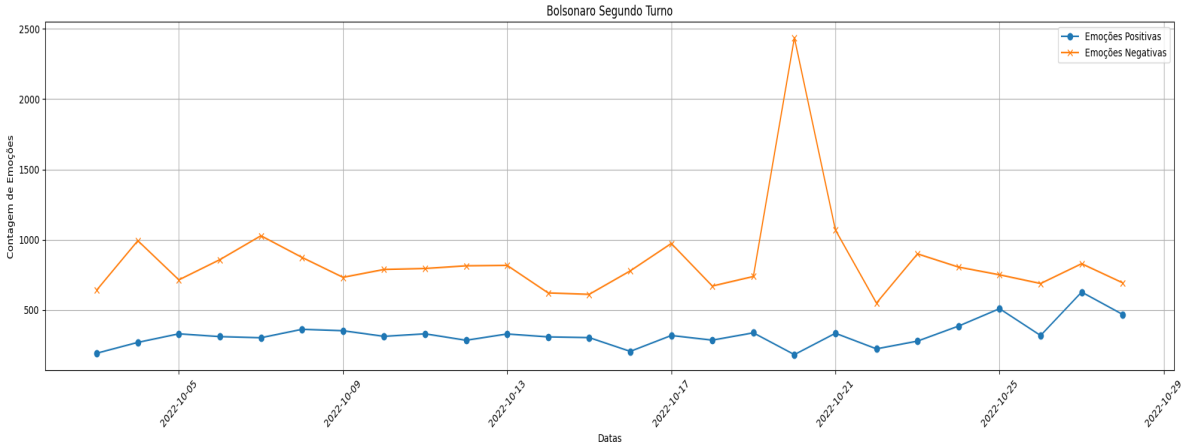

Figura 6. Segundo turno Bolsonaro.
Fonte: Autores.

Analisando os gráficos das Figuras 3 a 6, podemos observar claramente que o volume de *posts* negativos é maior para ambos os candidatos ao longo de todo o período eleitoral. Além disso, existem picos mais elevados, especialmente no caso dos *posts* negativos, provavelmente em resposta a algum evento relacionado ao candidato ocorrido no mesmo dia.

[13] https://www.em.com.br/app/noticia/politica/2022/08/18/interna_politica,1387507/youtuber-divulga-video-do-momento-da-reacao-de-bolsonaro.shtml

Candidato	Métrica	Turno			
		1º		2º	
		positivo	negativo	positivo	negativo
Lula	Média	314,7	555,6	291.6	524,7
	Variância	6118,1	5436,7	5310.7	7127,9
	Desvio Padrão	78,2	73,7	72.8	84,4
Bolsonaro	Média	329,6	873,2	325	852.5
	Variância	6776,5	40145,3	8648	116740.7
	Desvio Padrão	82,3	200,3	92.9	341.6

A média de *posts*, tanto positivos quanto negativos, para ambos os candidatos, Lula e Bolsonaro, é relativamente próxima. No entanto, Bolsonaro tem uma média maior de *posts* negativos em comparação com Lula, tanto no primeiro quanto no segundo turno. Para ambos os candidatos, a média de *posts* positivos e negativos diminuiu ligeiramente do primeiro para o segundo turno. Isso pode sugerir que, à medida que as eleições avançavam, o volume geral de discussão sobre os candidatos no X diminuiu um pouco.

A variância, que indica o grau de dispersão dos dados, é maior para Bolsonaro, especialmente quando se trata de *posts* negativos. Isso sugere que houve mais inconsistência nos sentimentos expressos em relação a Bolsonaro do que a Lula. As opiniões negativas sobre Bolsonaro foram mais variadas e menos previsíveis do que as opiniões negativas sobre Lula, ou seja, as opiniões negativas sobre Bolsonaro foram influenciadas por uma gama mais ampla de fatores ou eventos.

O desvio padrão, que também é uma medida de dispersão, segue um padrão semelhante à variância. Para Bolsonaro, o desvio padrão é significativamente maior para *posts* negativos, indicando uma maior variação nos sentimentos expressos. A variância e o desvio padrão dos *posts* negativos de Bolsonaro aumentaram significativamente, indicando que a dispersão das opiniões negativas sobre ele aumentou no segundo turno.

Para ambos os candidatos, a média, a variância e o desvio padrão dos *posts* positivos e negativos não mudaram significativamente do primeiro para o segundo turno. Isso indica que a opinião pública em relação a cada candidato se manteve relativamente estável ao longo das eleições. A análise desses dados sugere que, embora ambos os candidatos tenham recebido uma quantidade considerável de *feedback* positivo e negativo, a opinião pública em relação a Bolsonaro foi mais variada e inconsistente, especialmente em termos de *feedback* negativo.

6. Conclusão

Este estudo teve como objetivo classificar o sentimento das postagens da rede social X, relacionadas aos candidatos à presidência, Bolsonaro e Lula. As postagens coletadas foram processadas e submetidas a um modelo de treinamento que utiliza aprendizado de máquina supervisionado para classificar as opiniões coletadas de forma não binária, isto é, "positiva", "negativa" ou "neutra". Ao aplicar o classificador *Naive Bayes*, a maioria das postagens foram classificadas como neutras. Diante disso, decidimos ajustar nossa abordagem, utilizando apenas os resultados positivos e negativos. Essa mudança proporcionou uma compreensão mais clara e precisa dos sentimentos expressos pelos usuários.

Após uma revisão, constatamos que os resultados são bastante satisfatórios, com métricas como precisão, acurácia e *recall* superando os 90% no conjunto de treinamento de teste, o que reforça a confiabilidade de nossas análises. Ambos os candidatos receberam mais comentários negativos do que positivos, indicando uma tendência predominantemente negativa nos *posts* coletados.

Para trabalhos futuros, planeja-se utilizar outras técnicas de classificação. Uma possibilidade, é a utilização de análise em outras redes sociais além do X, a fim de obter opiniões dos usuários de forma mais abrangente. Também consideramos a aplicação de outras técnincas de classificação como o *k-means* para classificar os dados. Planeja-se também identificar padrões emocionais. Outro ponto de interesse para pesquisa é a análise da homofilia política entre os dados coletados, investigando como grupos com opiniões semelhantes influenciam uns aos outros. Além disso, planejamos examinar a relação entre os sentimentos expressos nas redes sociais e os resultados eleitorais, com o objetivo de antecipar tendências de comportamento dos eleitores.

Referências

Andleeb, S., Ahmed, R., Ahmed, Z., and Kanwal, M. (2019). Identification and classification of cybercrimes using text mining technique. In *2019 International Conference on Frontiers of Information Technology (FIT)*, pages 227–232. IEEE.

Benevenuto, F., Ribeiro, F., and Araújo, M. (2015). Curso de curta duração no brazilian symposium on multimedia and the web (webmedia).

Bragança, F. and Braga, B. (2023). Inteligência artificial e o impulsionamento de conteúdos nas redes sociais. Consultor Jurídico. Acessado em: 05 de maio de 2024. Disponível em: `https://www.conjur.com.br/2023-mar-31/braganca e-braga-ia-impulsionamento-redes-sociais/`.

Caetano, J. A. C., Lima, H. S. L., dos Santos Santos, M. F., and Marques-Neto, H. T. M.-N. (2017). Utilizando análise de sentimentos para definição da homofilia política dos usuários do twitter durante a eleição presidencial americana de 2016. In *Congresso da Sociedade Brasileira de Computação-CSBC*.

Chandak, A. (2022). Sentiment analysis on tweets in the 2020 us presidential election. *Journal of High School Science*, pages 1–18.

Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U., and Janjua, S. H. (2021). Sentiment analysis of before and after elections: Twitter data of u.s. election 2020. *Electronics (Switzerland)*, 10.

Dutra, D. A. M. and Francisco, E. R. (2018). Text mining: análise de sentimentos nas eleições 2018. In *Congresso Transformação Digital 2018*.

He, W., Zhang, W., Tian, X., Tao, R., and Akula, V. (2019). Identifying customer knowledge on social media through data analytics. *Journal of Enterprise Information Management*, 32(1):152–169.

Kumar, S., Kar, A. K., and Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, 1(1):100008.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Massetti, L. and Fernandes, V. (2017). Mineração de opinião aplicada ao cenário político. In *VI Encontro Acadêmico de Computação - UFMA*, pages 1–6, São Luis, Maranhão.

Matos, F., Magalhães, L., and Rocha Souza, R. (2020). Recuperação e classificação de sentimentos de usuários do twitter em período eleitoral. *Informação & Informação*, 25(1):92–114.

Pereira, J. G. (2019). Análise de sentimentos da populaçao brasileira em relaçao a eleiçao presidencial de 2018 através da rede social twitter. B.S. thesis, Universidade Federal do Rio Grande do Norte, Caicó, RN.

Queiroz, G. G. and Almeida, L. (2020). Uma metodologia de análise de sentimentos dos candidatos as eleições presidenciais de 2018 no twitter. *Revista de Engenharia e Pesquisa Aplicada*, 5(1):21–30.

Redhu, S., Srivastava, S., Bansal, B., and Gupta, G. (2018). Sentiment analysis using text mining: a review. *International Journal on Data Science and Technology*, 4(2):49–53.

Souza, M. F. d. L., Targa, C. N., and Silva, C. A. (2022). Mineração de opinião aplicada a postagens do twitter sobre o ensino remoto emergencial em institutos federais. *Revista de Sistemas e Computação-RSC*, 12(2).

Statista (2022). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. `https://www.statista.com/statistics/871513/worldwide-data-created/`. Accessed: 2023-02-01.

Suter, J., Nogueira, R., Croda, L., and Anderle, D. (2019). Desenvolvimento de um sistema de análise de sentimento utilizando técnicas de data warehousing. In *Simpósio Brasileiro de Banco de Dados - SBBD*, pages 295–300, Fortaleza, CE.