CONIC SEMESP

24º Congresso Nacional de Iniciação Científica

TÍTULO: MINERAÇÃO DE OPINIÃO NA REDE SOCIAL X SOBRE AS ELEIÇÕES PRESIDENCIAIS DE 2022

CATEGORIA: CONCLUÍDO

ÁREA: CIÊNCIAS EXATAS E DA TERRA

SUBÁREA: Estatística

INSTITUIÇÃO: INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS - IFMG

AUTOR(ES): BRUNO VICTOR VASCONCELOS

ORIENTADOR(ES): CARLOS ALEXANDRE SILVA, CRISTIANE TARGA

CATEGORIA CONCLUÍDO

1. RESUMO

As redes sociais, comumente presentes no cotidiano das pessoas, tem sido bastante utilizada 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*.

2. 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 parte da rotina de muitas pessoas, ganharam ainda mais importância durante o período eleitoral de 2022 no Brasil como espaços de interação e comunicação. Plataformas como Facebook, Instagram e *X* (anteriormente *Twitter*) são amplamente utilizadas e servem como recursos para diversos trabalhos sobre mineração de opinião, oferecendo ambientes onde vários tipos de conteúdo podem ser criados e compartilhados. O *X*, em particular, proporciona um ambiente onde distintos assuntos ganham relevância nacional e mundial, gerando uma enorme quantidade de dados para coleta e análise. Segundo (Bragança & Braga, 2023), o Brasil conta com aproximadamente 171,5 milhões de usuários ativos em redes

sociais, representando 79,9% da população brasileira, um crescimento de 14,3% ou 21 milhões de usuários entre 2021 e 2022[1].

O aumento de usuários em redes sociais tem impulsionado a geração e utilização de dados globalmente, com projeções indicando que esse volume ultrapassará 180 zettabytes em 2025 (Statista, 2022). Neste cenário, informações textuais são amplamente utilizadas para diversos fins, como gestão de serviços (Kumar, Kar, & Ilavarasan, 2021), prevenção de crimes cibernéticos (Andleeb, Ahmed, Ahmed, & Kanwal, 2019) e caracterização de clientes (He, Zhang, Tian, Tao, & Akula, 2019). (Liu, 2010) classifica essas informações em fatos (expressões objetivas) e opiniões (expressões subjetivas descrevendo sentimentos ou avaliações). Por exemplo, "*O presidente foi eleito com mais de 50% dos votos*" é um fato, enquanto "*O presidente está fazendo um ótimo trabalho*" é uma opinião. Este trabalho foca especificamente na análise de opiniões, destacando a importância desta distinção no contexto da pesquisa.

Este estudo analisou *posts* relacionados às eleições presidenciais de 2022, selecionando publicações que incluíam "Lula" ou "Bolsonaro" dentro das datas do primeiro e segundo turno, limitando-se aos 5 mil primeiros *posts* diários para um controle eficiente da base de dados. Utilizando o algoritmo de *Naive Bayes* após a limpeza dos dados, a análise revelou que a maioria dos *posts* expressava opiniões negativas sobre os candidatos. Esta abordagem, conhecida como mineração de texto ou análise de opinião, visa quantificar o valor emocional em textos, proporcionando uma compreensão mais profunda das opiniões expressas (Redhu, Srivastava, Bansal, & Gupta, 2018). Quando obtidas de redes sociais e devidamente analisadas, essas informações podem contribuir para a compreensão, explicação e até predição de complexos fenômenos sociais (Benevenuto, Ribeiro, & Araújo, 2015).

3. OBJETIVOS

Objetivo Geral

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

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

2022, Bolsonaro e Lula, em positivos, negativos ou neutros, a fim de compreender e avaliar a percepção pública em relação a esses políticos.

Objetivos específicos

- Construção de uma base de dados
- Realização de pré-processamento dos dados e utilização do algoritmo de *Naive Bayes* para classificar os *posts*.

4. 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*[2] para realizar a coleta dos *posts* aplicando filtros de busca específicos. Foram realizadas duas coletas, uma durante o primeiro turno e outra 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* com intuito de limpar e analisar os dados que seriam utilizados para treinamento do modelo. Esses *posts* contêm informações gerais a respeito do governo de Minas Gerais em 2017. Além disso, os *posts* desta base foram

2 https://github.com/JustAnotherArchivist/snscrape

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 utilizada 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.

5. DESENVOLVIMENTO

Foi realizada a coleta dos dados para a construção da base utilizando um algoritmo desenvolvido em Python com a biblioteca *snscrape*[3]. Em seguida, passamos a base para o algoritmo criado no *Google Colab*[4] 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*[5]. 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*[6] 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 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.

---

3 https://pypi.org/project/snscrape/ 6https://colab.google/

4 https://colab.google/

5 https://github.com/stacktecnologias/stack-repo/blob/master/Tweets Mg.csv

6 https://scikit-learn.org/stable/modules/naive bayes.html

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. 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[7] corridos. 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 à 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.

6. RESULTADOS

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

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.

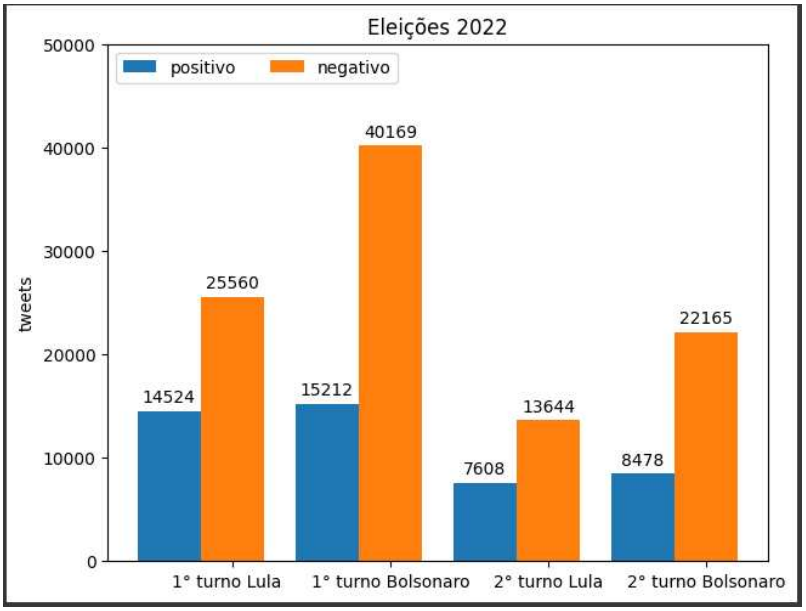

Figura 2. Comparativo de tweets positivos e negativos do 1º e 2º turno dos candidatos presidenciáveis. **Fonte:** Autores.

A análise inicial de um grande conjunto de *posts* nas redes sociais revelou um desafio: a grande quantidade de *posts* classificados como neutros dificultava a interpretação dos dados, impactando as Figuras 3 a 6. Para superar esse obstáculo, ajustamos nossa abordagem, focando apenas em *posts* positivos e negativos. Essa mudança resultou em uma compreensão mais clara e precisa dos sentimentos expressos pelos usuários.

Os resultados indicam que, embora ambos os candidatos tenham recebido mais menções negativas do que positivas, o candidato Bolsonaro foi alvo de um número significativamente maior de *posts* negativos em relação aos positivos. Essa discrepância sugere uma rejeição mais expressiva ao candidato por parte dos usuários das redes sociais.

As Figuras 3 a 6, agora com foco em *posts* positivos e negativos, apresentam de forma clara a evolução da percepção pública em relação aos candidatos. As linhas em cada gráfico representam o número diário de *posts* positivos e negativos durante

a campanha, destacando a dinâmica dos sentimentos ao longo do tempo. Essas figuras permitem identificar padrões e tendências na aprovação de cada candidato, fornecendo *insights* valiosos sobre a percepção do público durante o processo eleitoral.

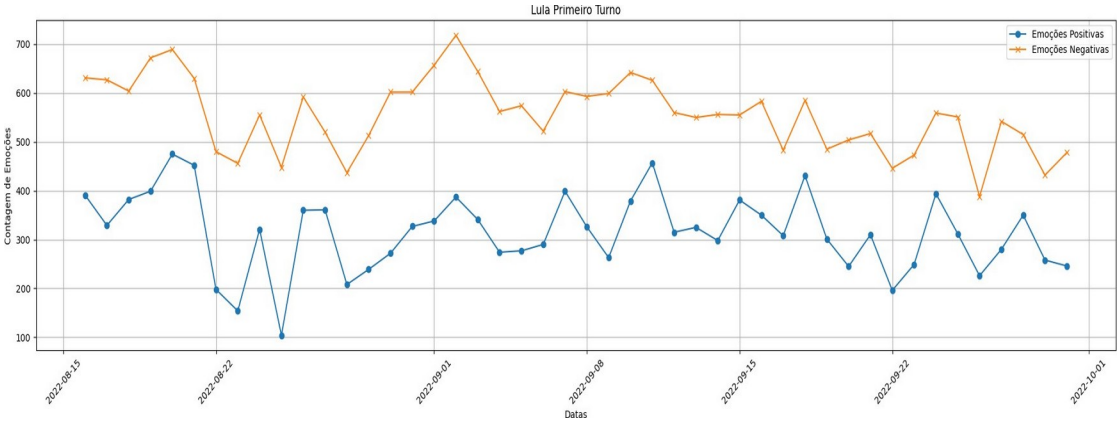

Figura 3. Primeiro turno Lula. **Fonte**: Autores.

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, houve uma diminuição tanto no número de *posts* positivos quanto negativos. Da mesma forma, 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[8], que revelou que Lula havia perdido dois pontos percentuais nas intenções de voto. Da mesma forma que a Figura 3 (referente ao 1º turno de Lula), a Figura 4 (relativa ao 2º 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.

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

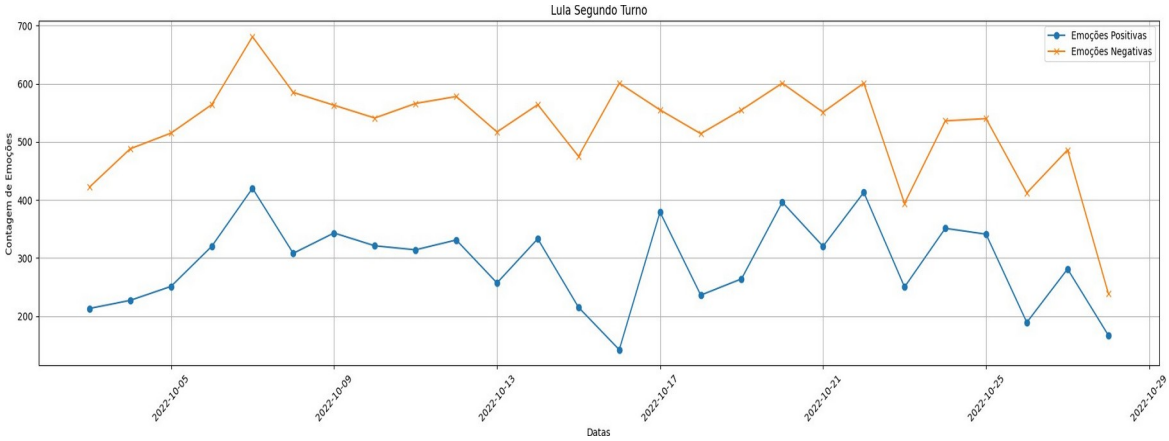

Figura 4. Segundo turno Lula. **Fonte:** Autores.

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[9]. 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.

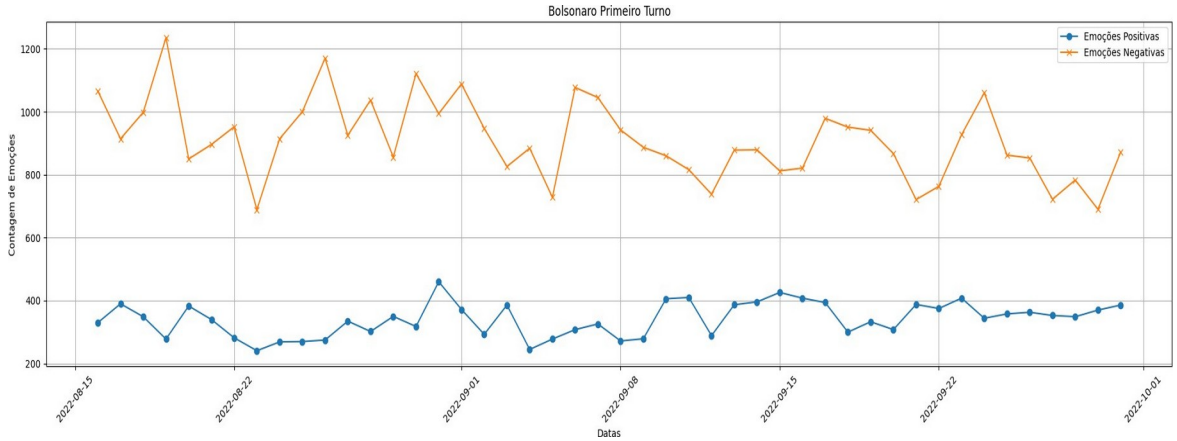

Figura 5. Primeiro turno Bolsonaro. **Fonte:** Autores.

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

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[10].

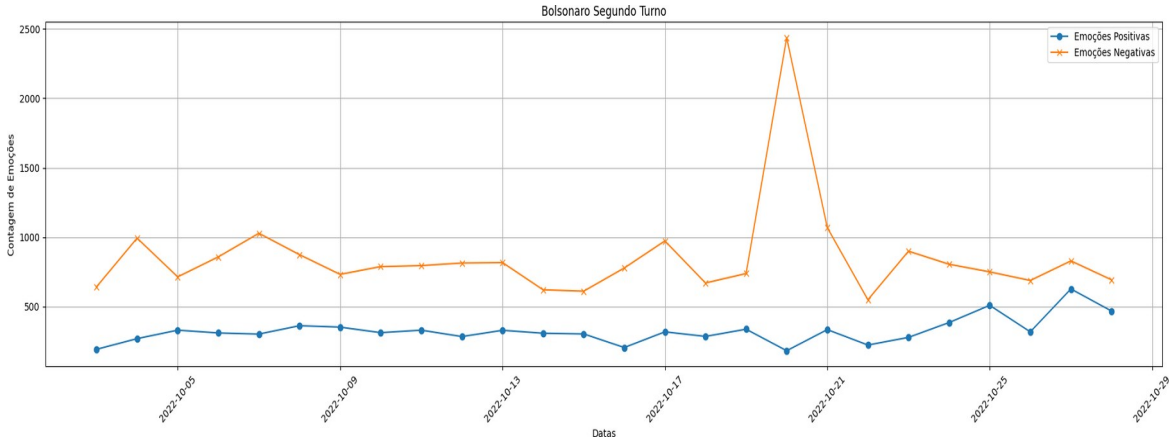

Figura 6. Segundo turno Bolsonaro. **Fonte:** Autores.

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.

7. CONSIDERAÇÕES FINAIS

Este estudo classificou o sentimento das postagens na rede social *X* relacionadas aos candidatos presidenciais Bolsonaro e Lula, utilizando aprendizado de máquina supervisionado. Inicialmente, a maioria das postagens foi classificada como neutra pelo classificador *Naive Bayes*, levando a uma abordagem ajustada que considerou apenas resultados positivos e negativos. As métricas de precisão, acurácia e recall superaram 90% no conjunto de treinamento de teste, indicando resultados

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

satisfatórios. Ambos os candidatos receberam mais comentários negativos do que positivos.

Para trabalhos futuros, planeja-se utilizar outras técnicas de classificação, como *k-means*, e expandir a análise para outras redes sociais. Pretende-se também identificar padrões emocionais, analisar a homofilia política entre os dados coletados e examinar a relação entre os sentimentos expressos nas redes sociais e os resultados eleitorais, visando antecipar tendências de comportamento dos eleitores.

8. FONTES CONSULTADAS

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

Benevenuto, F., Ribeiro, F., & Araújo, M. (Outubro de 2015). Curso de curta duração no Webmedia. *Brazilian Symposium on Multimedia and the Web (Webmedia*.

Bragança, F., & Braga, B. (Março de 2023). *Inteligência artificial e o impulsionamento de conteúdos nas redes sociais.* Fonte: Consultor Jurídico: https://www.conjur.com.br/2023-mar-31/bragancae-braga-ia-impulsionamento-redes-sociais/

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

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 14*, pp. 1-14.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, pp. 627--666.

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

Statista. (2022). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025.* Acesso em 1 de Fevereiro de 2023, disponível em https://www.statista.com/statistics/871513/worldwide-data-created/