Aprendizado de Máquina e Análise de Sentimento em Redes Sociais: Um Estudo de Caso Usando as Eleições Presidenciais em 2022

Bruno La Gatta Oliveira
Centro Federal de Educação Tecnológica
Campus Leopoldina (CEFET-MG)
Leopoldina, Brasil
brunolagatta55525@gmail.com

Luan Soares Oliveira
Centro Federal de Educação Tecnológica
Campus Leopoldina (CEFET-MG)
Leopoldina, Brasil
laun@cefetmg.com

***Abstract*—This study reports on the process of building and evaluating training models using textual data extracted from social networks in the Brazilian electoral runoff scenario. The objective is an evaluation of the relevance and quality that publicly exposed opinions have for political censuses in comparison to traditional lagged electoral surveys. The texts are extracted, processed, classified, and evaluated from the point of view of the feelings contained, which directly reflects on the voting intentions expressed as rejection or approval directed at the candidate. A relationship was established between the results obtained and the actual results from popular research sources in the country.**

***Keywords*—Data mining; Classification algorithms; Election polls.**

***Resumo*—O presente estudo relata o processo de construção e avaliação de modelos de treinamento utilizando de dados textuais extraídos de redes sociais no cenário de segundo turno das eleições presidenciais brasileiras de 2022. O objetivo é uma avaliação da relevância e qualidade que as opiniões expostas publicamente possuem para censos políticos em comparação a pesquisas eleitorais tradicionais defasadas. Os textos são extraídos, processados, classificados e avaliados do ponto de vista dos sentimentos contidos, o que reflete diretamente nas intenções de voto expressas como rejeição ou aprovação direcionado ao candidato. Estabeleceu-se uma relação entre os resultados obtidos com os resultados reais e de fontes de pesquisa populares no país.**

***Palavras-chave*—Mineração de dados; Algoritmos de classificação; Pesquisas eleitorais.**

I. Introdução

Com o avanço tecnológico, a análise de sentimentos, que tem como objetivo determinar a intensidade de sentimentos e a polaridade das frases capturadas da internet [1] em redes sociais, tem se tornado uma área de pesquisa em constante evolução. Os dados gerados pelos usuários nas redes sociais têm sido utilizados como uma fonte valiosa de informação para diferentes áreas como nos trabalhos de Jadhav [2], aplicado na área econômica e *marketing*, como também na política, em um trabalho semelhante a este, de Oliveira et al. [3], que faz um estudo comparativo entre o uso de análise de sentimentos e as pesquisas de intenção de votos tradicionais. O uso de técnicas de aprendizado de máquina para análise de sentimentos em redes sociais têm se mostrado uma abordagem promissora para entender as opiniões e atitudes dos usuários, como também é apresentado por Oliveira et al. [3], onde os resultados apresentados obtiveram percentual de variação de 1% a 8% em relação aos métodos tradicionais.

Da forma como é aplicada atualmente, a pesquisa eleitoral apresenta problemas de abrangência e especificidade. O uso de questionários e entrevistas possui algumas desvantagens. Sobre o primeiro, pode-se citar a impossibilidade do respondente de tirar dúvidas sobre as perguntas no momento de respondê-las. Já com relação às entrevistas, algumas desvantagens são: baixa abrangência, necessidade de treinar os entrevistadores e elevado gasto de tempo para entrevistar muitos usuários [4].

Alguns estudos já foram realizados utilizando a análise de sentimentos em redes sociais em contextos políticos, como o estudo de Caetano [5] relacionado a homofilia política nas eleições americanas utilizando o Twitter, como também o já citado estudo de Oliveira et al. [3]. Esses estudos têm mostrado que as redes sociais podem ser uma ferramenta poderosa para entender a opinião pública e a percepção dos eleitores. Além disso, o alcance das redes sociais têm chegado aos mais diversos extratos da sociedade, possibilitando uma coleta de dados abrangente.

De fato, estamos cada vez mais conectados às redes sociais, e isso tem gerado um volume enorme de dados que pode ser explorado para entender melhor as opiniões e atitudes das pessoas. Segundo uma pesquisa realizada em 2021 pela Hootsuite e *We Are Social*, existem 4,20 bilhões de usuários de mídia social em todo o mundo. Esse número

cresceu 490 milhões nos últimos 12 meses, gerando um crescimento ano a ano de mais de 13%. No mesmo estudo, é apontado que cerca de 70,3% da população brasileira são membros ativos de redes sociais [6].

Com base nisso, propõe-se neste artigo um estudo de caso sobre a aplicação de aprendizado de máquina e análise de sentimentos em redes sociais durante as eleições presidenciais de 2022. O objetivo é desenvolver uma aplicação que extraia textos publicados no Twitter, processe-os e classifique os sentimentos dos usuários em relação aos candidatos à presidência. Acreditamos que os resultados obtidos podem fornecer informações valiosas para os candidatos, às equipes de campanha e aos eleitores.

O presente artigo será apresentado a partir desta introdução, seguida do referencial teórico, da metodologia aplicada e da apresentação dos resultados obtidos, finalizando com as considerações finais, trabalhos futuros e referências.

II. Referencial teórico

Para Boiy e Moens [7] a análise de sentimento pode ser definida como a extração de sentimentos de uma fonte não estruturada, como textos, imagens ou áudios. Também defendido por Boiy e Moens, a análise de sentimento ganhou foco devido a percepção da importância da opinião das pessoas em relação a diversos tópicos sendo a internet, uma excelente fonte dessas informações. A polaridade das frases, como define Rosa [8], representa as características positivas, negativas ou neutras da frase. Segundo Rosa, "os sentimentos expressam o grau de intensidade positiva ou negativa de uma frase, possuindo uma escala que pode variar".

O conceito de mineração de dados para Gomes et al. [9] tem como objetivo categorizar ou encontrar padrões em grandes volumes de dados em diferentes mídias. Já segundo Costa et al. [10], a mineração de dados pode ser definida como uma das etapas de um processo superior conhecido como descoberta de conhecimento em bases de dados, ou KDD (*Knowledge Discovery in Databases*). Costa et al. [10] relatam que o KDD consiste primeiramente em uma etapa de pré-processamento dos dados, baseado na preparação de dados, abrangendo mecanismos para captação, organização e tratamento dos dados. A seguir, é feita efetivamente a mineração e, após essa, é feito o pós-processamento dos dados obtidos na fase de mineração.

Mineração de texto pode ser definida como "Processo de Descoberta de Conhecimento, que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras." [11]. Segundo Morais e Ambrósio, esse processo envolve a aplicação de algoritmos computacionais que processam os textos e identificam em seu conteúdo informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado.

Os dados coletados passam por uma fase de pré-processamento, com objetivo de solucionar problemas nos dados, visualizá-los ou alterar sua estrutura, além de melhorar os resultados na etapa de extração de conhecimento [12]. Dentro dessa fase, técnicas como remoção de *stopwords* e *steeming* podem ser utilizadas [13]. Na vetorização, conceito da transformação dos textos pré-processados em vetores, podem ser utilizados algoritmos como o *Bag of Words* [14] e o TF-IDF (*Term-Frequency-Inverse Document Frequency*) [15]. A importância da aplicação destas técnicas se dá pela necessidade de uniformizar os dados, reduzindo variações linguísticas e transformando os dados em uma estrutura matemática, que servirá de entrada para os algoritmos de aprendizado de máquina.

Aprendizado de máquina pode ser definido como "a capacidade de um computador "aprender" a partir de um conjunto de dados e, com isso, fazer previsões ou classificações, sem que tenha sido previamente programado." [16]. Para Monard e Baranauskas [17], no aprendizado supervisionado é fornecido para o sistema a ser treinado uma base de exemplos classificados a serem reproduzidos. O intuito é mapear todos os dados associados a suas classificações para construir o modelo. A qualidade dos resultados dessa classificação, por sua vez, pode ser mensurada a partir de valores numéricos, chamados de medidas de desempenho.

Medidas de desempenho são parâmetros universais aplicados à análise dos resultados de bases de treinamento com o objetivo de reconhecer a capacidade classificatória das mesmas, podendo ser apresentados em diferentes tipos de valores [18]. Dentre os quais pode-se citar acurácia/erro, medida F, AUC, dentre outros [19]. Essas medidas são comumente agrupadas e manipuladas em uma estrutura conhecida por Matriz de Confusão, que relaciona em duas dimensões as medidas verdadeiras e as médias preditas para diferentes classes (como positivo e negativo) dos modelos estudados [17].

III. Metodologia

O presente estudo de caso aplicado ao segundo turno das eleições brasileiras de 2022 foi subdividido em etapas seguindo o modelo de KDD iniciado em seleção, pré-processamento, transformação, mineração de dados e avaliação.

A primeira etapa consistiu no estudo da API (*Application Programming Interface*) da plataforma escolhida, o twitter, utilizada em sua versão v2. A utilização do twitter se deve ao fato de ser uma rede social fortemente textual, contendo uma quantidade de textos de opinião maior que as demais redes sociais populares. A linguagem escolhida para o desenvolvimento da aplicação foi o Python, devido à sua grande aplicabilidade nas áreas de inteligência artificial e tratamento de dados, devido à sua gama de bibliotecas, dando especial foco ao Pandas, usado para a manipulação e pré-processamento dos textos, e Scikit-learn, que, dentre seus métodos, possui todos os algoritmos de treinamento utilizados neste estudo. Teve-se início o

processo de codificação da conexão com a API para a extração dos dados, tratamento e armazenamento. Os *tweets* foram armazenados em arquivo tipo CSV (valores separados por vírgulas), tratados utilizando a biblioteca pandas. Para os atributos requisitados, foram armazenados data de criação da conta, localização, parâmetros públicos (seguidores, usuários seguidos, número de *tweets* e listas públicas na qual o usuário é membro) do usuário e texto e data de publicação do *tweet* em questão.

Os algoritmos MultinomialNB (*Multinomial Naive Bayes*), algoritmo baseado nos estudos de Thomas Bayes e KNN (*K-Nearest Neighbors*) foram objetos de estudo por Devika et al. [20]. Em seu estudo "*Sentiment Analysis: A Comparative Study On Different Approaches*", é feita uma análise entre diferentes aproximações e algoritmos utilizados para o processo de análise de sentimentos, apontando pontos positivos e negativos em cada método. *Random Forest* (RF) por sua vez já foi motivo de estudo por Karthika et al. [21] no qual foi feito uma análise comparativa de sua aplicação no âmbito da análise de sentimentos com outro algoritmo de classificação, o SVM (*Support Vector Machine*). Em um estudo de classificação de questões de exames escolares, baseado em níveis cognitivos que retrata corretamente o nível de dificuldade das questões, realizado por Sulaiman et al. [22], foram utilizados os métodos de classificação pelos modelos de treinamento Naive Bayes, KNN e o Linear SVC (*Linear Support Vector Classification*), sendo o Linear SVC o que apresentou melhores resultados. O trabalho de Ricci [16] é uma análise de sentimentos referente à reforma da previdência em 2019 no Brasil, utilizando como fonte o Twitter. Nele, o mesmo utiliza de diferentes algoritmos de classificação, dentre eles a Regressão Logística, utilizando a mesma biblioteca do presente estudo. Os métodos de *Random Forest* e KNN possuem, em relação ao seu processo classificatório, um parâmetro configurável atrelado diretamente a seu desempenho. Para o *Random Forest*, o parâmetro ***n*** diz respeito ao número de árvores de decisão que serão geradas em cima do vetor de dados. Já o KNN, possui o parâmetro ***k***, relacionado ao número de "vizinhos" próximos de determinada instância no vetor a ser analisada. Foi feita a variação dos mesmos em relação a seus valores predefinidos em suas configurações documentadas conforme [23], sendo ***n*** variado entre 50, 100 e 150, e ***k*** variado entre 3, 5 e 7. Os algoritmos foram aplicados sobre os dados de acordo com as seguintes configurações de parâmetros:

- MultinomialNB: configuração padrão definida pela documentação;
- RF: Variação do ***n*** (*n_estimators*) entre 50, 100 e 150 para cada teste utilizando este algoritmo. Valores padrão para os demais parâmetros;
- KNN: Variação do ***k*** (*n_estimators*) entre 3, 5 e 7 para cada teste utilizando este algoritmo. Valores padrão para os demais parâmetros;
- Linear SVC: Os parâmetros *random_state*, *tol*, dual, *max_iter* foram alterados para os valores 0, 1-e5, *False* e 120000 respectivamente. Os demais parâmetros estão configurados com valores padrão;
- Regressão Logística: Para este, apenas os parâmetros de *random_state*, *dual*, *max_iter* foram alterados para 0, *False* e 120000 respectivamente. Demais parâmetros seguem os valores padrão previstos na documentação.

Os testes preliminares foram executados utilizando a base de dados de análises fílmicas do IMDB (*Internet Movie Database*) já classificadas. A base contém 49045 análises em português de diversos filmes, divididas entre positivas e negativas. Para os testes, foi selecionada uma amostra de 5000 análises para cada tipo de sentimento. A fim de verificar também a eficácia dos métodos de pré-processamento, os 10000 dados analisados foram fragmentados em quatro bases distintas e, em cada uma delas, aplicou-se diferentes técnicas. As quatro passaram por um processamento de linguagem natural nas quais foram removidas as *stopwords*, links e outras palavras desnecessárias contidas nos textos. Na primeira base foi aplicada apenas a técnica de *Bag of Words*. Na segunda, foi aplicada a mesma técnica, porém, os dados passaram pelo processamento do *stemming*. Na terceira, foi utilizado TF-IDF. Na última, foi aplicado tanto o TF-IDF quanto o *stemming*. Destas, a subdivisão para os algoritmos de se baseou na proporção de 70% de treinamento, ou seja, utilizadas para treinar a base e 30% para testes, para verificar o quão efetivo foi o treinamento. Foram executados cinco testes com dados selecionados aleatoriamente dentro da divisão estabelecida. Cada algoritmo analisou cada uma das bases cinco vezes, gerando 180 matrizes de confusão. As matrizes de confusão geradas seguem o modelo fornecido pela função *classification_report* da biblioteca scikitlearn. O principal atributo analisado foi a acurácia, para compreensão de qual algoritmo e qual base gerou os melhores resultados de classificação.

A terceira etapa consistiu na extração dos *tweets*, processamento e classificação manual. A classificação manual é importante para servir de base de treinamento e teste para os algoritmos. Foram classificados manualmente 672 *tweets* seguindo dois modelos classificatórios: um binário (positivo e negativo) e um de multiclasses (positivo, negativo e neutro). Para o primeiro, 365 eram positivos e 307 negativos. No segundo utilizando dos mesmos *tweets*, foram classificados 216 positivos, 255 negativos e 201 neutros. O critério classificatório se baseou no sentimento geral e intenções do texto apresentado. Da API do Twitter, foram extraídos textos durante o período de 07/10/2022 a 30/10/2022, totalizando 17260 tweets. Para os treinamentos dos algoritmos em cima destes dados, foram separadas amostras de 300/300 e 200/200/200 para cada modelo.

IV. Resultados

Diante dos procedimentos descritos na metodologia, a análise inicial dos algoritmos mencionados utilizando as bases de testes processadas do IMBD, foram obtidos os seguintes valores médios de acurácia extraídos da matriz de confusão representados na Tabela 1:

TABELA I
VALORES DE ACURÁCIA DE CLASSIFICAÇÃO (EM PORCENTAGEM) OBTIDOS DOS ALGORITMOS EM DIFERENTES BASES

Algoritmo	*BoW*	*BoW e Stemming*	*TF-IDF*	*TF-IDF e Stemming*
Multinomial NB	74	89	78	90
RF (n = 50)	69	83	66	82
RF (n = 100)	69	85	66	83
RF (n = 150)	68	86	64	84
KNN (k = 3)	50	62	74	84
KNN (k = 5)	51	63	75	85
KNN (k = 7)	50	63	74	86
Linear SVC	68	68	64	65
RL	73	88	76	89

Pelos resultados, nota-se que as bases de dados que continham tanto o uso do TF-IDF quanto de stemming obtiveram melhores valores de acurácia em mais da metade dos casos. Além disso, o algoritmo *Naive Bayes* retornou o maior valor de acurácia. Diante dessa análise, os dados extraídos do twitter passaram pelos mesmos passos de pré-processamento, e foram testados com todos os algoritmos levantados.

Durante o processo de classificação manual dos *tweets* extraídos, feito após o pré-processamento dos textos, foram observadas duas características no conjunto de dados que poderiam prejudicar os resultados. Primeiramente, havia textos repetidos de fontes diferentes, o que levaria a classificações repetidas e poderiam atrapalhar o treinamento dos modelos. Segundo, existiam alguns casos de múltiplos *tweets* por usuário, o que levava em consideração a opinião do mesmo mais de uma vez. Portanto, dos 17259 *tweets* extraídos, 2641 foram removidos devido às razões citadas, resultando em 14.618 analisados. Utilizando dos modelos classificatórios já descritos, foram obtidos os valores de acurácia para cada algoritmo, dispostos na Tabela 2:

TABELA 2
VALORES DE ACURÁCIA DE CLASSIFICAÇÃO (EM PORCENTAGEM) OBTIDOS DOS ALGORITMOS NAS BASES CLASSIFICADAS DO TWITTER PARA CADA MODELO CLASSIFICATÓRIO

Algoritmo	Binário	Multiclasse
MultinomialNB	82	55
RF (n = 59)	78	54
KNN (k = 1)	75	54
Linear SVC	77	60
Regressão Logística	83	60

Foi observado que a classificação de multiclasse apresentou resultados piores, levando a adoção do método binário de classificação, juntamente com o algoritmo de Regressão Logística para a reprodução em escala da classificação da base. Os dados, então, foram classificados pelo modelo treinado.

O assunto de cada *tweet* foi dividido entre os que faziam referência aos candidatos Luiz Inácio Lula da Silva, Jair Messias Bolsonaro e àqueles comentários considerados neutros, isto é, que não faziam alusão direta a nenhum dos candidatos. Por fim, foi adicionada a classificação dos dois tipos de sentimentos: positivo e negativo. A Figura 1 mostra os resultados obtidos.

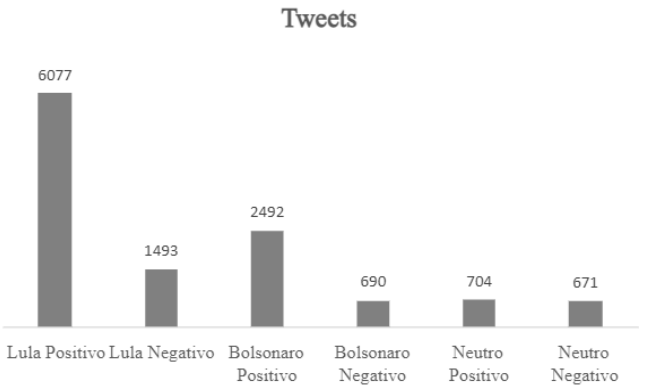

Fig. 1. Gráfico de *tweets* classificados.

Diante dos dados, foi feita uma análise de proporcionalidade do total de *tweets* para candidato, separados por índice de aprovação e rejeição, numa analogia aos sentimentos positivos e negativos. A tabela 3 mostra os resultados obtidos:

TABELA 3
VALORES DE ACURÁCIA DE CLASSIFICAÇÃO (EM PORCENTAGEM) OBTIDOS DOS ALGORITMOS NAS BASES CLASSIFICADAS DO TWITTER PARA CADA MODELO CLASSIFICATÓRIO

Candidato	Aprovação	Rejeição	Total
Luiz Inácio Lula da Silva	6077 (80,3%)	1493 (19,7%)	7570 (100%)
Jair Messias Bolsonaro	2492 (78,3%)	690 (21,7%)	3182 (100%)
Comentários Neutros	704 (51,2%)	671 (48,8%)	1375 (100%)

Segundo os resultados das eleições divulgados pelo TSE (Tribunal Superior Eleitoral), o candidato Luiz Inácio Lula da Silva venceu com 50,9% dos votos, contra 49,1% de Jair Bolsonaro. As pesquisas eleitorais na véspera das eleições pelo instituto DATAFOLHA (instituto de pesquisas do Grupo Folha) apresentaram um placar de 52% a 48% [24], enquanto pelo IPEC (Inteligência em Pesquisa e Consultoria Estratégica) apresentou um placar de 54% a 46% [25]. Os resultados mostram um cenário de equilíbrio entre os candidatos, o que converge com os dados de aprovação/rejeição em relação aos candidatos encontrados no trabalho. Os números absolutos de *tweets* em relação a cada candidato podem ter pouca relevância, devido a variação do perfil dos eleitores.

V. Considerações finais e trabalhos futuros

Em uma análise qualitativa dos dados obtidos, algumas considerações podem ser feitas. Foram encontrados vários *tweets* repetidos, o que pode ser explicado pela possível existência de *bots* (contas ilegítimas) que propagam os mesmos textos repetidamente. Apesar dos textos repetidos terem sido retirados da base de dados final, é possível a construção de um algoritmo mais inteligente para detecção de *bots*, que leve em conta a análise de alguns dados sobre a conta (número de seguidores, data de criação, número de curtidas, etc), visto que esta camada extra de filtragem pode melhorar a confiabilidade dos dados obtidos.

Outro fator a ser levado em consideração é o fato de que é difícil garantir que a amostra coletada representa um extrato proporcional da população. Isso se explica pela ausência na base de dados de características dos eleitores julgadas importantes nesse tipo de pesquisa, como escolaridade e renda. Tais dados não são fornecidos pelo Twitter, visto que não é necessário que os usuários os forneçam para criar uma conta e publicar *tweets*. Essas informações são relevantes para identificar picos de popularidade em diferentes grupos de eleitores, o que pode indicar que determinado grupo tem mais interesse nas propostas de certos candidatos.

Além destes, as características peculiares da linguagem escrita são também um desafio. É complexo para algoritmos interpretar figuras de linguagem, como ironia e sarcasmo, tão presentes no cenário político atual e, principalmente, em redes sociais. Outro fator complicador é a utilização de diferentes nomes para se referir ao mesmo candidato, como nomes pejorativos e apelidos disseminados entre os usuários de redes sociais. Isso dificulta tanto a análise do sentimento contido nos textos quanto a identificação do sujeito alvo das falas, além de prejudicar a definição de um dicionário de palavras-chave para busca e extração dos textos através da API.

No que tange a proporcionalidade dos resultados obtidos, é natural comparar a quantidade bruta de dados referentes a cada candidato. Isso pode ser resultado de diferentes fatores: a rede social pode ser mais popular para determinado grupo de eleitores, pode haver concentração de campanhas de marketing por parte dos partidos em diferentes mídias, entre outros fatores. Apesar disso, as análises mais aprofundadas discutidas neste artigo, que convergem com as estimativas dos institutos de pesquisa e com o resultado final das eleições, foram feitas com base em dados proporcionais, e não em quantidade bruta.

Por fim, em um escopo geral, se evidenciam dois outros fatores prejudiciais à qualidade dos resultados: As limitações da API em extração de *tweets*, com permissão para extrair no máximo 100 *tweets* por palavra chave em um dado intervalo de tempo; e a existência de poucas bases de teste já classificadas em português para contextos políticos. Analisando essa segunda como fator motivador para trabalhos futuros, já se encontra em desenvolvimento uma base de dados para modelos de treinamento em português no âmbito de *benchmarks* e avaliações comerciais de produtos, que podem possuir termos próximos aos cenários políticos.

Agradecimentos

Agradeço à instituição Centro Federal de Educação Tecnológica de Minas Gerais, por todo o apoio e suporte durante a construção desta pesquisa.

Referências

[1] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up?: Sentiment Classification Using Machine Learning Techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10, 79-86. https://doi.org/10.3115/1118693.1118704

[2] Jadhav, R., & M. S., W. (2017). Survey: Sentiment Analysis of Twitter Data for Stock Market Prediction. IJARCCE, 6(3), 558–562. https://doi.org/10.17148/ijarcce.2017.63129

[3] Oliveira, D. J. S., Bermejo, P. H. d. S., & dos Santos, P. A. (2016). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. Journal of Information Technology & Politics, 14(1), 34–45. https://doi.org/10.1080/19331681.2016.1214094

[4] Barbosa, S., & Silva, B. (2010). Interação humano-computador. Elsevier Brasil.

[5] Caetano, J., Lima, H., dos Santos, M., & Marques-Neto, H. (2017). Utilizando Análise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016. In Anais do VI Brazilian Workshop on Social Network Analysis and Mining. Porto Alegre: SBC. doi:10.5753/brasnam.2017.3246

[6] Digital 2021 - We Are Social UK. (s.d.). We Are Social UK. https://wearesocial.com/uk/blog/2021/01/digital-2021-uk/

[7] Boiy, E., & Moens, M.-F. (2008). A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval, 12(5), 526–558. https://doi.org/10.1007/s10791-008-9070-z

[8] Rosa, R. L. (2015). Análise de sentimentos e afetividade de textos extraídos das redes sociais. Tese de Doutorado, Escola Politécnica, Universidade de São Paulo, São Paulo. doi:10.11606/T.3.2016.tde-19072016-115713. Recuperado em 2023-07-03, de www.teses.usp.br

[9] Gomes, J., Pimenta, R., & Schneider, M. (2019). no campo ENANCIB. Acesso em https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/546/517

[10] Costa, E., Baker, R. S., Amorim, L., Magalhães, J., & Marinho, T. (2012). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em Informática na Educação, 1(1), 1-29.

[11] Morais, E. A. M., Ambrósio, A. P. L. (2007). Mineração de Textos (Relatório Técnico - INF_005/07). https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf

[12] Batista, G. E. A. P. A. (2003). Pré-processamento de dados em aprendizado de máquina supervisionado. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. doi:10.11606/T.55.2003.tde-06102003-160219. Recuperado em 2023-07-03, de www.teses.usp.br

[13] Martins, C. A. (2003). Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo.

[14] Matsubara, E. T., Martins, C. A., & Monard, M. C. (2003). Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words.

[15] Manning, C. D. (2008). Introduction to Information Retrieval. Cambridge University Press.

[16] Ricci, R. D. (2020). Análise de sentimentos no Twitter sobre a Reforma da Previdência no ano de 2019. Trabalho de Conclusão de Curso, Graduação em Estatística, Universidade Federal de Uberlândia, Uberlândia. Recuperado em 2023-07-03, de https://repositorio.ufu.br/handle/123456789/30900

[17] Monard M. C., Baranauskas A. J. (2003). Conceitos sobre Aprendizado de Máquina. In Sistemas inteligentes Fundamentos e aplicações (p. 89–114). Manole Ltda.

[18] Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters, 30(1), 27–38.

[19] Souza, B. F. (2010). Meta-aprendizagem aplicada à classificação de dados de expressão gênica. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. doi:10.11606/T.55.2010.tde-04012011-142551. Recuperado em 2023-07-03, de www.teses.usp.br

[20] DEVIKA, M. D.; SUNITHA, C.; GANESH, Amal. Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science, v. 87, p. 44-49, 2016.

[21] Karthika P., Murugeswari R. and Manoranjithem R. (2019). Sentiment Analysis of Social Media Network Using Random Forest Algorithm. IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 1-5.

[22] Sulaiman, S., Wahid, R. A., Hanee Ariffin, A., & Zalina Zulkifli, C. (2002). Question Classification Based on Cognitive Levels using Linear SVC. Test Engineering and Management, 83, 6463–6470.

[23] SCIKIT-LEARN: machine learning in Python — scikit-learn 1.3.1 documentation. Disponível em: https://scikit-learn.org/stable/. Acesso em: 11 out. 2023.

[24] Folha. (2022, 31 de outubro). Na véspera da eleição, Lula tem 52%, e Bolsonaro, 48%. Folha de São Paulo. https://datafolha.folha.uol.com.br/eleicoes/2022/10/na-vespera-da-eleicao-lula-tem-52-e-bolsonaro-48.shtml

[25] G1. (2022, 29 de outubro). Ipec: Lula tem 54% dos votos válidos no 2º turno, e Bolsonaro, 46%. G1. https://g1.globo.com/politica/eleicoes/2022/pesquisa-eleitoral/noticia/2022/10/29/ipec-lula-tem-54percent-dos-votos-validos-2o-turno-e-bolsonaro-46percent.ghtml