O Boletim de Conjuntura (BOCA) publica ensaios, artigos de revisão, artigos teóricos e empíricos, resenhas e vídeos relacionados às temáticas de políticas públicas.

O periódico tem como escopo a publicação de trabalhos inéditos e originais, nacionais ou internacionais que versem sobre Políticas Públicas, resultantes de pesquisas científicas e reflexões teóricas e empíricas.

Esta revista oferece acesso livre imediato ao seu conteúdo, seguindo o princípio de que disponibilizar gratuitamente o conhecimento científico ao público proporciona maior democratizacão mundial do conhecimento.

BOLETIM DE CONJUNTURA

BOCA

Ano VI | Volume 19 | Nº 55 | Boa Vista | 2024

http://www.ioles.com.br/boca

ISSN: 2675-1488

https://doi.org/10.5281/zenodo.13372607

BOLETIM DE CONJUNTURA

www.ioles.com.br/boca

IDENTIFICAÇÃO DE DISCURSO DE ÓDIO NAS ELEIÇÕES MUNICIPAIS DE 2020

Brenda Paetzoldt Silva[1]

Fernando Santos[2]

Resumo

O discurso de ódio está se tornando frequente nas redes sociais, e evidências indicam que ele aumenta nos anos em que há eleição. É frequentemente pauta de discussões sobre políticas públicas para responsabilizar seus autores e mitigar seus efeitos sem ferir a liberdade de expressão dos indivíduos. Neste sentido, a inteligência artificial pode ser uma aliada, auxiliando na identificação do discurso de ódio. Este estudo apresenta a construção e avaliação de um classificador *Naïve Bayes* para identificar discurso de ódio em publicações na rede social X relacionadas a candidatos que disputaram o segundo turno das eleições municipais de 2020. O trabalho adotou metodologia quantitativa. Para construção e avaliação do classificador *Naïve Bayes* proposto foi adotado o processo CRISP-DM, consolidado na área de ciência de dados. Os dados utilizados para treinamento do classificador foram levantados de maneira exploratória no Kaggle, um repositório online de dados para treinamento de modelos preditivos. Já os dados utilizados para avaliação do modelo foram coletados a partir de um levantamento de publicações realizadas na rede social X. O desempenho do classificador foi avaliado de forma quantitativa, a partir de métricas estatísticas de assertividade. De modo geral, o classificador *Naïve Bayes* proposto obteve acurácia média de 72,38% no conjunto de publicações coletadas para avaliação. A diferença de desempenho do classificador proposto em relação a *Perspective API*, uma ferramenta online para identificação de discurso de ódio adotada como referência no presente estudo, ficou abaixo de 9,12% em todos os candidatos considerados. A partir destes resultados verifica-se que o classificador *Naïve Bayes* proposto foi capaz de identificar a presença de discurso de ódio em publicações da rede social X relacionadas a candidatos que disputaram o segundo turno das eleições municipais de 2020.

Palavras-chave: Discurso de Ódio; Eleições; Inteligência Artificial; Rede Social.

Abstract

Hate speech has become increasingly prevalent on social media, with evidence suggesting that it intensifies during election years. It is frequently a topic of discussion in public policy debates, with the aim of holding perpetrators accountable and mitigating its effects without infringing upon individual freedom of expression. In this context, artificial intelligence can be a valuable tool to aid in the identification of hate speech. This study presents the construction and evaluation of a Naïve Bayes classifier to identify hate speech in publications on social media platform X related to candidates who contested the second round of the 2020 municipal elections. A quantitative methodology was employed. The CRISP-DM process, well-established in the field of data science, was adopted for the construction and evaluation of the proposed Naïve Bayes classifier. The data used to train the classifier was collected through an exploratory search on Kaggle, an online repository of data for training predictive models. The data used to evaluate the model was collected from a survey of publications on social media platform X. The classifier's performance was evaluated quantitatively using statistical accuracy metrics. Overall, the proposed Naïve Bayes classifier achieved an average accuracy of 72.38% on the dataset collected for evaluation. The performance difference between the proposed classifier and Perspective API, an online tool for hate speech identification adopted as a reference in this study, was less than 9.12% for all candidates considered. These results demonstrate that the proposed Naïve Bayes classifier was capable of identifying the presence of hate speech in publications on social media platform X related to candidates who contested the second round of the 2020 municipal elections.

Keywords: Artificial Intelligence; Election; Hate Speech; Social Network.

[1] Bacharela em Engenharia de Software pela Universidade do Estado de Santa Catarina (UDESC). E-mail: brendapaetzoldt@gmail.com

[2] Professor da Universidade do Estado de Santa Catarina (UDESC). Doutor em Ciência da Computação. E-mail: fernando.santos@udesc.br

INTRODUÇÃO

Redes sociais estão cada vez mais presentes no quotidiano das pessoas. Através delas, usuários publicam suas rotinas diárias, pensamentos e opiniões. Enquanto muitas destas publicações são inofensivas, algumas podem conter discurso de ódio. De acordo com a Organização das Nações Unidas, discurso de ódio é qualquer tipo de comunicação (falada, escrita ou comportamental) que ataca ou usa linguagem pejorativa ou discriminatória com referência a uma pessoa ou grupo. O ataque é baseado em fatores de identidade da pessoa, como por exemplo, religião, etnia, nacionalidade, raça, cor e gênero. Este tipo de discurso pode gerar intolerância e ódio.

Em eleições para cargos políticos, candidatos têm utilizado contas pessoais e partidárias como veículos de comunicação e captação de votos. Contudo, de acordo com a Safernet Brasil (2022), as eleições são um gatilho para discursos de ódio na internet. Em períodos eleitorais, a presença de discurso de ódio em publicações pode ter seu efeito amplificado dado o alcance das publicações de candidatos, especialmente a cargos majoritários. Ainda segundo a Safernet Brasil, as denúncias de crimes de ódio aumentam em anos de eleição, sendo que nos primeiros meses de 2022, 23.947 denúncias foram recebidas, quantidade 67,5% superior ao mesmo período de 2021.

Uma das recomendações do Ministério dos Direitos Humanos e da Cidadania para enfrentamento ao discurso de ódio é a responsabilização dos divulgadores. Contudo, para responsabilizá-los é necessário primeiramente identificar as publicações que apresentam discurso de ódio. A identificação manual, realizada por moderadores humanos, é inviável para atender ao grande volume de conteúdo produzido nas redes sociais. Neste sentido, a Organização das Nações Unidas para Educação, Ciência e Cultura (UNESCO) aponta a relevância de se considerar técnicas de aprendizagem de máquina para automatizar a identificação do discurso de ódio. Além disso, a UNESCO também enfatiza que muitas iniciativas para identificação automatizada do discurso de ódio estão limitadas ao idioma inglês. Isto ampara e justifica pesquisas como a apresentada neste trabalho, que investigam a identificação automática de discurso de ódio em redes sociais considerando o idioma português do Brasil.

A inteligência artificial, por meio de suas técnicas de aprendizagem de máquina e mineração de dados, tem sido aliada no estudo e avaliação de políticas públicas. Dentre as técnicas de aprendizagem de máquina muito utilizadas para construção de ferramentas capazes de classificar textos e análise de sentimentos está a *Naïve Bayes*. Uma *Naïve Bayes* é um classificador probabilístico que, após treinado, é capaz de determinar a probabilidade de um texto ser de determinada classe (e.g., ser discurso de ódio) a partir das evidências observadas (e.g., palavras presentes no texto). Apesar da simplicidade, classificadores *Naïve Bayes* têm proporcionado bons resultados em aplicações reais, principalmente

classificação de documentos, requerendo reduzida quantidade de dados para treinamento, e sendo especialmente adequada quando a classificação é baseada em muitos atributos. Neste sentido, os classificadores *Naïve Bayes* e seu sucesso reportado em aplicações de classificação de texto caracterizam o marco conceitual que sustenta o presente estudo.

O objetivo desta pesquisa é construir um classificador *Naïve Bayes* para identificar discurso de ódio em publicações da rede social X no idioma português do Brasil, e avaliar seu desempenho considerando publicações da rede social X relacionadas a candidatos que disputaram o segundo turno das eleições municipais de 2020 no Brasil.

O trabalho adotou metodologia quantitativa. A questão de pesquisa considerada é: como identificar, de forma automatizada, o discurso de ódio em publicações da rede social X relacionadas às eleições municipais de 2020? A hipótese considerada neste trabalho é que a identificação do discurso de ódio nessas publicações pode ser realizada através de um classificador *Naïve Bayes*, haja visto o sucesso reportado para estes classificadores quando aplicados para classificação de textos.

Para construção do classificador *Naïve Bayes*, este trabalho adotou o processo denominado *CRoss Industry Standard Process for Data Mining* (CRISP-DM), já consolidado na área de ciência de dados. O classificador foi treinado utilizando um conjunto de dados obtido de forma exploratória através do repositório Kaggle, e avaliado sobre este conjunto de dados aplicando-se métricas estatísticas de assertividade. Nesta avaliação o classificador *Naïve Bayes* proposto obteve acurácia média de 72,38% na identificação de discurso de ódio. Já para avaliação da capacidade do classificador identificar discurso de ódio em publicações relacionadas às eleições municipais, foram coletadas publicações da rede social X relacionadas a candidatos de municípios onde houve segundo turno nas eleições municipais de 2020. Os resultados obtidos sobre estas publicações foi comparado com a ferramenta *online Perspective* API, que identifica discurso de ódio em textos e foi adotada como referência no presente estudo. Nesta comparação, a diferença de desempenho entre o classificador proposto e a *Perspective* API ficou abaixo de 9,12% em todos os candidatos considerados, evidenciando sua capacidade de identificar discurso de ódio nessas publicações.

Este texto está organizado em cinco seções, sendo a primeira essa introdução. A seção 2 apresenta a fundamentação teórica sobre discurso de ódio, o classificador *Naïve Bayes*, e discute o estado da arte a partir de trabalhos relacionados. A seção 3 descreve a metodologia adotada para construção do classificador *Naïve Bayes*, e a seção 4 apresenta os resultados e discussões do trabalho a partir da identificação de discurso de ódio nas publicações do X. Por fim, a seção 5 expõe as considerações finais e perspectivas de trabalhos futuros.

DISCURSO DE ÓDIO

Ao longo do tempo, tem-se percebido a importância de que todos sejam livres para expressar suas opiniões. No Brasil, por exemplo, pode-se notar uma grande preocupação e legislações que asseguram o direito de que todas as pessoas possam dizer o que pensam e opinar sobre diferentes temas. Este direito é garantido principalmente no Art. 220. da Constituição Federal, que afirma que "a manifestação do pensamento, a criação, a expressão e a informação, sob qualquer forma, processo ou veículo não sofrerão qualquer restrição, observado o disposto nesta Constituição" (BRASIL, 1988).

A liberdade de expressão é essencial para a democracia, sendo protegida por muitas leis e documentos internacionais. Segundo Gargarella (2011): "sem liberdade de expressão, não há democracia. Ela ocupa o centro nevrálgico de uma estrutura democrática". Todavia, existe uma linha tênue entre o que pode ser considerado direito e o que é um crime. O direito à liberdade de expressão não confere imunidade ao indivíduo para permitir comportamentos difamatórios ou caluniosos. É neste contexto que se situa o discurso de ódio. Ele está relacionado com o crime de ódio pois o principal objetivo de seus atos é incitar o ódio e a desarmonia entre as pessoas. Os crimes de ódio são muitas vezes motivados por animosidade pessoal e por convicção política (PEREIRA; MEDEIROS; COUTINHO, 2020).

O discurso de ódio se manifesta por meio de uma violência psicológica que também está incluída como crime, principalmente quando isso leva ao dano emocional e diminuição da autoestima por meio de ameaças, constrangimento, humilhação, manipulação, isolamento, vigilância constante, perseguição e insulto (BRASIL, 2014). No contexto eleitoral, o discurso de ódio surge uma vez que:

> Os eleitores tendem a se rivalizar uns com outros [por meio de] discursos de ódio [...], repetidamente espalhados pelas novas tecnologias proporcionadas pelas redes sociais [...] Com ataques aos costumes, às instituições da democráticas, às liberdades individuais, à liberdade de imprensa, com discursos de ódio que se espalham por redes sociais e mentiras (*fakenews*) sobre as mais variadas questões -inclusive sobre as formas de votação -os líderes autocratas e seus séquitos minam as bases da democracia (ELIAS; BRASIL, 2024, p. 698; 713).

No projeto de lei 7582/2014, discurso de ódio é definido como "ofensa à vida, à integridade corporal, ou à saúde de outrem motivada por preconceito ou discriminação em razão de classe e origem social, condição de migrante, refugiado ou deslocado interno, orientação sexual, identidade e expressão de gênero, idade, religião, situação de rua e deficiência" (BRASIL, 2014).

As denúncias de crimes de ódio aumentam em anos de eleição (SILVA, 2018). Visando auxiliar na redução da ocorrência de crimes envolvendo discursos de ódio, a organização não governamental Safernet Brasil[3] oferece um serviço para recebimento de denúncias de crimes e violações contra direitos

humanos na internet, contando com procedimentos efetivos e transparentes para lidar com as denúncias. Além disso, contam com suporte governamental, parcerias com a iniciativa privada, autoridades policiais e judiciais. Ainda segundo a Safernet Brasil, as denúncias de crimes de ódio aumentam em anos de eleição, sendo que nos primeiros meses de 2022, 23.947 denúncias foram recebidas, quantidade 67,5% superior ao mesmo período de 2021 (SAFERNET, 2022). Ainda que as eleições presidenciais, devido à abrangência nacional, possam gerar maior engajamento nas redes sociais, o discurso de ódio também é percebido em eleições regionais para cargos municipais (BISPO, 2020).

CLASSIFICADOR NAÏVE BAYES

Uma tarefa de classificação consiste em identificar a qual categoria (ou rótulo, ou classe) certo objeto pertence, a partir de seus atributos (ou variáveis). Na tarefa de classificação, costuma-se considerar que a classe é o atributo alvo, cujo valor deseja-se determinar a partir dos demais atributos do objeto. A classificação pode ser realizada por métodos de aprendizagem de máquina que, após treinados, predizem a classe de objetos (FACELI *et al*., 2021). Um destes métodos é o *Naïve Bayes*.

Segundo Escovedo & Koshiyama (2020), o *Naïve Bayes* é um método muito utilizado por ser rápido computacionalmente e necessitar pequena quantidade de dados para treinamento, sendo particularmente adequado quando os objetos possuem grande número de atributos. Provost & Fawcett (2016) destacam que o *Naïve Bayes* possui um desempenho supreendentemente bom para classificação em muitas tarefas do mundo real. Conforme Escovedo & Koshiyama (2020), o método *Naïve Bayes* é muito utilizado em aplicações de classificação de textos e análise de sentimentos em redes sociais.

O método *Naïve Bayes* é chamado *naïve* (ingênuo) por supor independência condicional entre as variáveis, ou seja, ele desconsidera eventual correlação entre as variáveis (ESCOVEDO; KOSHIYAMA, 2020). É baseado no Teorema de Bayes, sendo, portanto, fundamentado no cálculo de probabilidades condicionais. A equação a seguir apresenta o Teorema de Bayes.

$$P(C = c|E) = \frac{P(E|C = c) * P(C = c)}{P(E)}$$

Os termos da equação do Teorema de Bayes são (PROVOST; FAWCETT, 2016):

- *C*: o atributo alvo, que identifica a classe do objeto a ser classificado.
- *c*: uma das possíveis classes (valores) do atributo alvo.
- *E*: uma evidência, ou seja, os valores observados para os demais atributos do objeto.

- *P(C = c)*: probabilidade *a priori* da classe, isto é, a probabilidade de que o atributo alvo *C* seja a classe *c* independente dos valores dos demais atributos, ou seja, independentemente de qualquer evidência observada.
- *P(E|C = c)*: probabilidade de observar a evidência *E*, ou seja, determinados valores nos atributos, quando a classe do atributo alvo é *C*.
- *P(E)*: probabilidade *a priori* da evidência *E*, ou seja, de se observar determinados valores nos atributos.
- *P(C = c|E)*: probabilidade condicional *a posteriori* de atribuir certa classe ao atributo alvo dada a evidência observada.

Observa-se que o *Naïve Bayes* resolve a tarefa de classificação determinando a probabilidade condicional *a posteriori* da classe, a partir dos atributos do objeto (evidência observada). As demais probabilidades envolvidas no cálculo são estimadas a partir do conjunto de dados utilizado no treinamento do *Naïve Bayes* (ESCOVEDO; KOSHIYAMA, 2020). Por se tratar de um método de aprendizagem supervisionado, é necessário que o conjunto de dados utilizado no treinamento esteja rotulado, ou seja, que cada texto contido neste conjunto já esteja identificado como contendo ou não discurso de ódio. Além disso, para evitar superadaptação aos dados utilizados, é prática adotar alguma estratégia de validação cruzada para treinamento do *Naïve Bayes*, selecionando uma fração dos dados para treino e outra para avaliação do desempenho do modelo preditivo.

Para utilizar aprendizagem de máquina na tarefa de classificação de texto, é necessário adotar um formato de representação propício ao método de classificação. No caso do *Naïve Bayes* um formato frequentemente adotado é a vetorização, que consiste na elaboração de uma matriz de frequência de termos (PROVOST; FAWCETT, 2016). Nessa matriz, cada linha representa um texto em particular (por exemplo, uma sentença, ou uma publicação), havendo uma coluna para cada palavra distinta existente no corpus linguístico. Cada célula dessa matriz contém um valor numérico que define a frequência de ocorrência da palavra (coluna) no texto (linha). A partir da matriz, o método *Naïve Bayes* consegue estimar as probabilidades necessárias para predizer a classe a partir do conjunto de dados de treinamento.

TRABALHOS RELACIONADOS

Os trabalhos relacionados a seguir foram encontrados através do *Google Schoolar* ao se utilizar palavras-chave relacionadas a classificação de discurso de ódio (*hate speech*, *discurso de ódio*, *modelos de classificações* e *eleições*).

Davidson *et al*. (2017) investigaram a detecção automática de discurso de ódio no X com o objetivo de verificar a viabilidade de separar a identificação de discurso de ódio de outros tipos de publicações ofensivas. Os autores justificam o trabalho devido ao fato de que abordagens existentes

baseadas em aprendizagem de máquina não foram efetivas em diferenciar estes dois tipos de publicações. Para o estudo, os autores coletaram publicações do X, no idioma inglês, e classificaram essas publicações como contendo discurso de ódio ou somente termos ofensivos. As publicações foram utilizadas para treinar um classificador de regressão linear multiclasse para distingui-las. Como resultados do estudo, os autores apontam que publicações com termos racistas ou homofóbicos tendem a ser classificadas como discurso de ódio, enquanto publicações com termos sexistas tendem a ser classificadas como ofensivas.

No trabalho desenvolvido por Salminen *et al*. (2020) foi proposta a coleta de 197.566 comentários em quatro plataformas: YouTube, Reddit, Wikipedia e X. Os dados foram coletados dos repositórios *Kaggle* e *Perspective* API, sendo 80% dos comentários classificados como não odiosos e o restante com a identificação de discurso de ódio. Após a seleção dos dados, foram experimentados diferentes métodos de classificação. Os resultados obtidos pelos classificadores foram comparados com os resultados originais dos conjuntos de dados previamente classificados, e os autores constataram que o método *XGBoost* proporcionou o melhor desempenho.

O trabalho desenvolvido por Ribeiro (2016) visa identificar a existência de sarcasmo, algo que é frequentemente adotado por usuários de redes sociais. O sarcasmo pode se tornar um problema na análise de sentimentos, já que o sentido ou a polaridade do texto pode ser totalmente invertida. Foram coletadas três mil mensagens do X na língua inglesa que continham *hashtags* relacionadas a sarcasmo (e.g.: *#irony*, *#lying* e *#notreally*). Parte desses dados foi submetida a uma análise humana sendo rotulado com o julgamento adequado acerca da classificação de cada mensagem. Os autores compararam os resultados obtidos por diferentes métodos classificadores, e reportaram que o *Naïve Bayes* obteve 97% de sucesso ao classificar corretamente a polaridade no conjunto de dados com a presença de *hashtags* de cunho sarcástico e 89% sucesso na base de dados sem *hashtags*.

Em um trabalho que buscou analisar a literatura relacionada a discurso de ódio, Tontodimamma *et al*. (2021) identificaram que a detecção e classificação automática de discurso de ódio utilizando técnicas de inteligência artificial está entre três tópicos mais pesquisados nos últimos 30 anos. Entre os achados do estudo, os autores apontam que o Brasil está entre os 10 países com maior quantidade de publicações relatando pesquisas sobre o debate entre discurso de ódio e liberdade de expressão.

Apesar disso, Yin e Zubiaga (2021) observam que as pesquisas envolvendo identificação de discurso de ódio consideram majoritariamente o idioma inglês. Isso é corroborado pelo trabalho de Alkomah e Ma (2022), que conduziram uma revisão da literatura sobre métodos e conjuntos de dados (*datasets*) para detecção de discurso de ódio em textos. Os *datasets* encontrados pelos autores são, majoritariamente, em inglês. Em outra revisão da literatura, Jahan e Oussalah (2023) buscaram técnicas

de processamento de linguagem natural que têm sido utilizadas para detectar automaticamente o discurso de ódio. Dentre as conclusões apontadas pelos autores, observa-se também a falta de estudos e experimentos com idiomas diferentes do inglês. A generalização de modelos preditivos interidiomas é uma alternativa apontada por Yin e Zubiaga (2021), mas que pode apresentar desafios dada a dissimilaridade entre linguagens e culturas.

No que diz respeito a estudos considerando o idioma português do Brasil, Vargas *et al*. (2022) avaliaram o desempenho de diferentes técnicas de aprendizagem de máquina em um *dataset* composto por publicações da rede social Instagram em português. Como resultado, reportaram que o melhor desempenho foi obtido com um classificador *Naïve Bayes*, obtendo precisão de 78% na identificação do discurso de ódio. No estudo de Oliveira *et al*. (2023), os autores avaliaram o desempenho do ChatGPT em identificar discurso de ódio em publicações da rede social X em português. Apesar do ChatGPT ter popularizado o potencial da inteligência artificial, os autores do estudo reportam que sua precisão na identificação de discurso de ódio foi de 74%.

Assim como os trabalhos supracitados, que investigam ou propõem o desenvolvimento de classificadores capazes de identificar discursos de ódio em comentários originários de redes sociais, o presente trabalho também propõe avaliar o desempenho de um classificador em textos da rede social X. O diferencial do presente trabalho está na origem do conjunto de dados utilizado para treinamento, e a comparação do classificador *Naïve Bayes* com a ferramenta *Perspective* API.

Cabe também destacar a baixa quantidade de trabalhos propondo a análise e classificação de textos na língua portuguesa com métodos da inteligência artificial, conforme evidenciado pelos trabalhos supracitados. Os conjuntos de dados pré-classificados disponíveis e as bibliotecas para implementação de aprendizagem de máquina que suportam a língua portuguesa são mais raras do que com a língua inglesa, o que contribui para evidenciar a relevância do presente trabalho.

METODOLOGIA

O presente trabalho se caracteriza como natureza primária, conduzida sob um procedimento técnico de estudo de caso exploratório, no qual se investigou em profundidade a tarefa de identificar discurso de ódio em publicações da rede social X relacionadas às eleições municipais de 2020 a partir da construção e avaliação um classificador *Naïve Bayes*.

A metodologia adotada para o trabalho é quantitativa, tendo como foco estudar o problema de identificar discurso de ódio em publicações da rede social X, de forma automatizada utilizando uma

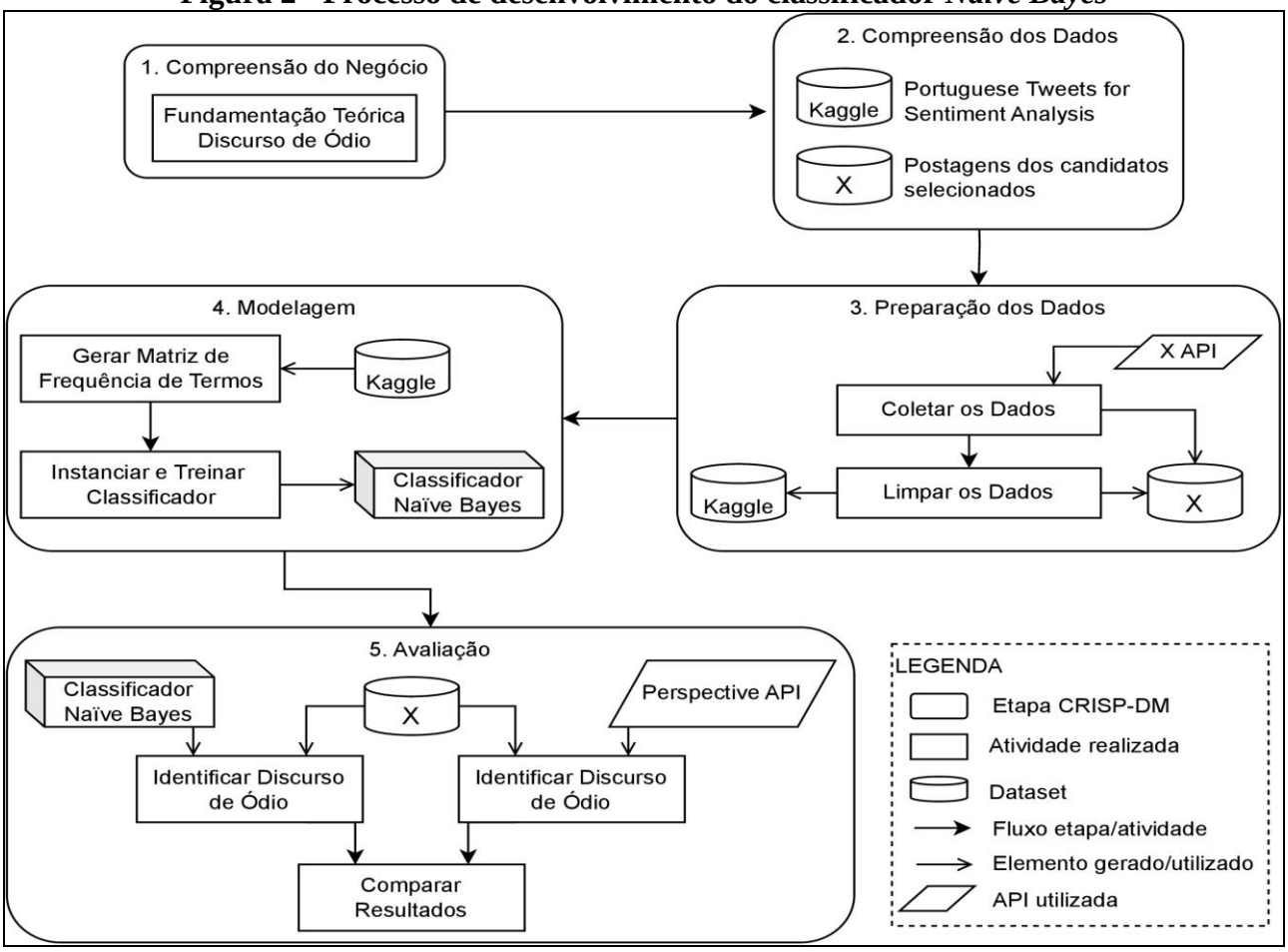
técnica inteligência artificial. Os resultados quantitativos obtidos possibilitam medir, numericamente, o desempenho do classificador *Naïve Bayes* construído com essa finalidade. Para obtenção destes resultados, o classificador *Naïve Bayes* construído foi empregado computacionalmente para identificar quais publicações coletadas da rede social X dos candidatos às eleições municipais de 2020 de determinadas cidades apresentam discurso de ódio. A escolha das cidades consideradas no estudo é detalhada posteriormente. A avaliação quantitativa do desempenho do classificador *Naïve Bayes* foi realizada a partir de métricas estatísticas para assertividade, em particular acurácia, *precision*, *recall* e *f1-score*.

Conforme previamente descrito, a construção de um classificador *Naïve Bayes* requer um conjunto de dados para treinamento. O procedimento adotado para levantamento destes dados foi exploratório, examinando-se um repositório de conjuntos de dados para uso em aprendizagem de máquina chamado Kaggle[4] e realizando-se pesquisas por palavras-chave “portuguese”, “tweets”, “hate speech”, “sentiment analysis”. Do ponto de vista da presente pesquisa, este conjunto de dados caracteriza-se como dados secundários. Os dados dos municípios considerados no presente estudo, também são considerados secundários e foram obtidos a partir do Instituto Brasileiro de Geografia e Estatística.

Sendo que o objetivo geral deste trabalho é construir e avaliar um classificador *Naïve Bayes* para identificação de discurso de ódio em publicações da rede social X, um processo consolidado em ciência de dados foi adotado para este fim. O processo, denominado *CRoss Industry Standard Process for Data Mining* (CRISP-DM) (WIRTH; HIPP, 2000), é formado por seis etapas: compreensão do negócio; compreensão dos dados; preparação dos dados; modelagem; avaliação; e implantação (CARVALHO *et al*., 2024). Apesar de sua origem remarcar aos anos 2000, o processo CRISP-DM segue sendo referência e adotado em projetos que envolvem mineração de dados e aprendizagem de máquina (RAWAT, 2023; ZAVALETA-SÁNCHEZ *et al*., 2024).

Neste trabalho foram adotadas as cinco primeiras etapas do processo CRISP-DM. A etapa de implantação não foi necessária tendo em vista que o classificador não foi implantado, ficando essa etapa como sugestão de trabalho futuro. A Figura 2, disposta na página seguinte, apresenta a visão geral do processo CRISP-DM conduzido neste trabalho. A seguir são descritas as atividades que foram realizadas nas etapas 1 a 3.

Figura 2 - Processo de desenvolvimento do classificador *Naïve Bayes*



Fonte: Elaboração própria.

Compreensão do negócio

A etapa de compreensão do negócio envolveu entender o contexto em que o classificador está inserido. Trata-se do contexto de identificação de discurso de ódio em publicações na rede social X relacionadas a candidatos que disputaram o segundo turno das eleições municipais de 2020 no Brasil. Nessa etapa buscou-se a fundamentação teórica necessária, e que foi previamente descrita.

Compreensão dos dados

Nesta etapa foi constatado que as publicações na rede social X se caracterizam como dados não estruturados, sendo formadas por palavras em idioma português do Brasil, sem qualquer classificação relacionada à presença de discurso de ódio.

Por se tratar de um método de aprendizagem supervisionada, o treinamento do *Naïve Bayes* requer um conjunto de dados (*dataset*) já classificado, ou seja, um conjunto de

publicações X que já estejam classificados como contendo ou não discurso de ódio. Para treinar o classificador *Naïve Bayes* proposto neste trabalho foi utilizado um *dataset* disponível no Kaggle (AUGUSTOP, 2018). Intitulado *Portuguese Tweets for Sentiment Analysis*, este conjunto de dados é composto ao todo por 800 mil publicações X já classificadas como contendo sentimento positivo, negativo ou neutro. Neste trabalho foi utilizado apenas o subconjunto de dados com publicações relacionadas à política brasileira, sendo formado por mais de 60 mil publicações no período de 01/08/2018 a 20/10/2018 e representado como *dataset* Kaggle na Figura 2. São 32.744 publicações (53%) políticas com sentimento positivo e 28.847 (47%) publicações políticas com sentimento negativo. Para cada publicação, o *dataset* contém um código identificador da publicação, o texto, a data de criação e o sentimento (positivo ou negativo). Os autores relatam ter utilizado o método de Go, Bhayani & Huang (2009) para classificação do sentimento.

Para avaliação do classificador *Naïve Bayes* foi elaborado um *dataset* contendo publicações dos candidatos selecionados para este trabalho. Chamado de *dataset* X na Figura 2, a sua criação envolveu a coleta de publicações na rede social, conforme descrito a seguir.

Preparação dos dados

Esta etapa foi dividida em duas subetapas: coletar dados e limpar dados. A etapa de coletar dados foi necessária apenas para construir o *dataset* X, pois o *dataset* Kaggle já estava disponível online, sendo necessário apenas realizar a limpeza dos dados.

1. Coleta de Publicações X

Para coletar os dados do *dataset* X foi utilizada a X API (X CORP, 2024), uma biblioteca que permite a seleção e coleta de dados via código. Um resumo em inglês deste trabalho foi submetido para a análise e aprovação da equipe da X API e, dez dias após, a solicitação para acesso nível *Academic Research* foi aprovada. Foi necessário para a coleta das publicações criadas no período de estudo que o acesso à API fosse nível acadêmico, já que em outros níveis só é permitido requisitar publicações feitas em no máximo sete dias antes da data da requisição.

A coleta de publicações foi implementada na linguagem *Python*. Através dos serviços disponibilizados pela X API são requisitadas as publicações desejadas. A requisição contém os parâmetros da consulta. Neste trabalho foram utilizados como parâmetros para consulta o destinatário da publicação (conta citada ou mencionada no texto) e o intervalo (datas e horas) das publicações

desejadas. Foi necessário informar o intervalo desejado pois a quantidade máxima de publicações retornadas na consulta é 500. Portanto a criação do *dataset* envolveu a repetição de diversas consultas, limitando o intervalo para períodos menores nos casos em que a quantidade de publicações retornadas era superior ao limite da X API. De cada publicação são salvos no *dataset* X o identificador (código) da publicação, a data e o texto da publicação.

O objetivo deste trabalho é identificar discurso de ódio em publicações relacionadas a candidatos que disputaram o segundo turno das eleições municipais de 2020. Para atender este objetivo, foram escolhidos cinco municípios que tiveram segundo turno das eleições e cujos candidatos possuíam conta na rede social X. Três municípios onde houve segundo turno foram selecionados por serem os mais populosos do país segundo o Instituto Brasileiro de Geografia e Estatística (2022). Os dois outros são municípios do estado de Santa Catarina onde houve segundo turno.

Foram coletadas, no dia 1 de junho de 2022, respostas das publicações e mensagens destinadas aos candidatos que possuem contas pessoais ou partidárias, ativas, realizadas no período eleitoral do segundo turno, de 16 de novembro de 2020 até 28 de novembro de 2020. Foram consideradas como contas ativas aquelas que contém ao menos uma publicação feita no período citado, e tenham recebido ao menos uma resposta para esta publicação. A tabela 1 apresenta a população, partido e conta X dos candidatos e respectivos municípios considerados neste trabalho. A tabela também mostra a quantidade de publicações que mencionam cada candidato. O resultado da coleta é um *dataset* com 16.852 publicações.

Tabela 1 - Municípios e candidatos considerados neste trabalho

Município	População*	Candidato	Partido	Conta no X	Posts
São Paulo (SP)	12.396.372	Bruno Covas	PSDB	@brunocovas	1209
		Guilherme Boulos	PSOL	@GuilhermeBoulos	5407
Rio de Janeiro (RJ)	6.775.561	Eduardo Paes	DEM	@eduardopaes	2187
		Marcelo Crivella	Republicanos	@MCrivella	3231
Fortaleza (CE)	2.703.391	Cap. Wagner	PROS	@capitao_wagner	1654
		Sarto	PDT	@sartoprefeito12	2545
Joinville (SC)	604.708	Adriano Silva	NOVO	@souadrianosilva	405
		Darci De Matos	PSD	@depdarcidematos	33
Blumenau (SC)	366.418	João P. Kleinubing	DEM	@jpkleinubing	99
		Mário Hildebrandt	PODE	@mariohildebrand	82

Fonte: Elaboração própria.
Nota: * Dados de Instituto Brasileiro de Geografia e Estatística (2022)

No caso do município de São Paulo, a conta do candidato Bruno Covas foi desabilitada da rede social X um ano após seu falecimento. Isso impediu que fossem requisitadas as publicações diretamente endereçadas à conta do candidato. Para contornar essa limitação, foram coletadas todas as publicações

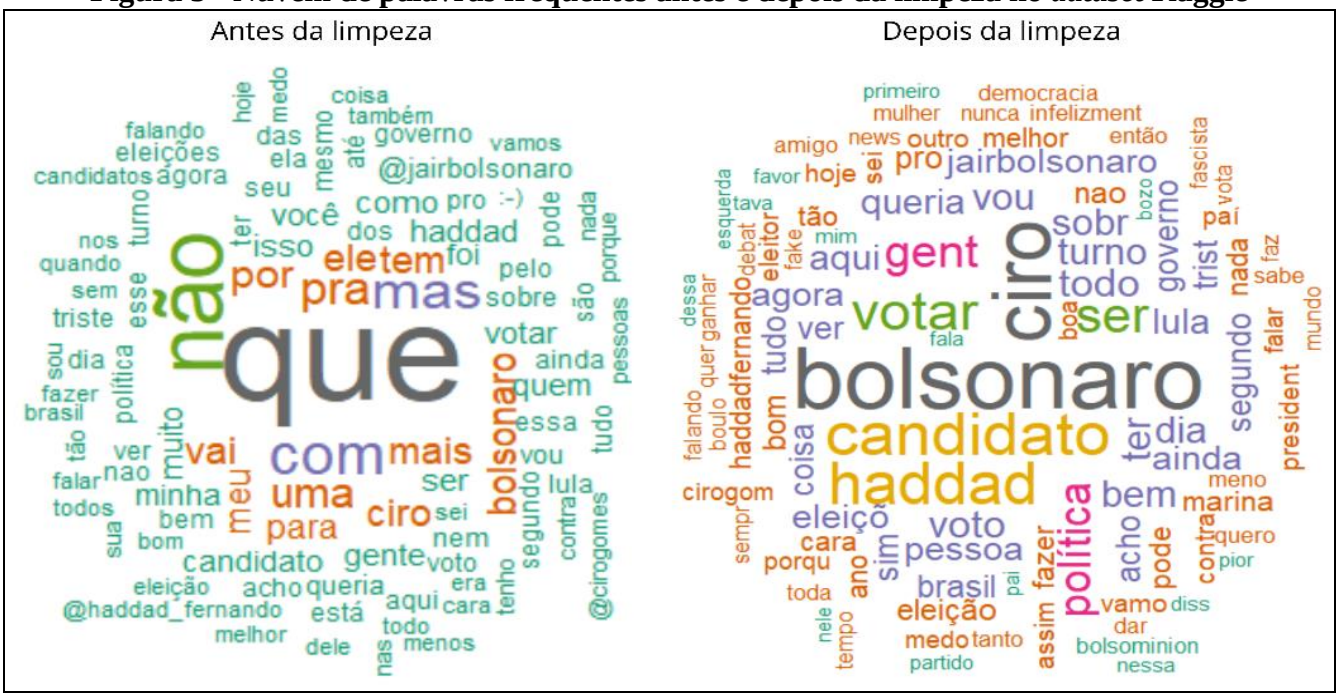
endereçadas ao candidato, mas sem que sua conta (sua "arroba") estivesse contida no texto. Isso explica a menor quantidade de publicações coletadas deste candidato em relação ao seu adversário. Apesar disso, cabe mencionar que isto não introduz qualquer viés no presente trabalho tendo em vista que este *dataset* não é utilizado para treinar o classificador.

2. *Limpeza dos Dados*

Para aplicar aprendizagem de máquina em classificação de texto, algumas operações devem ser realizadas para limpar os dados e descartar o que não possui valor semântico para a classificação (SILVA, 2016).

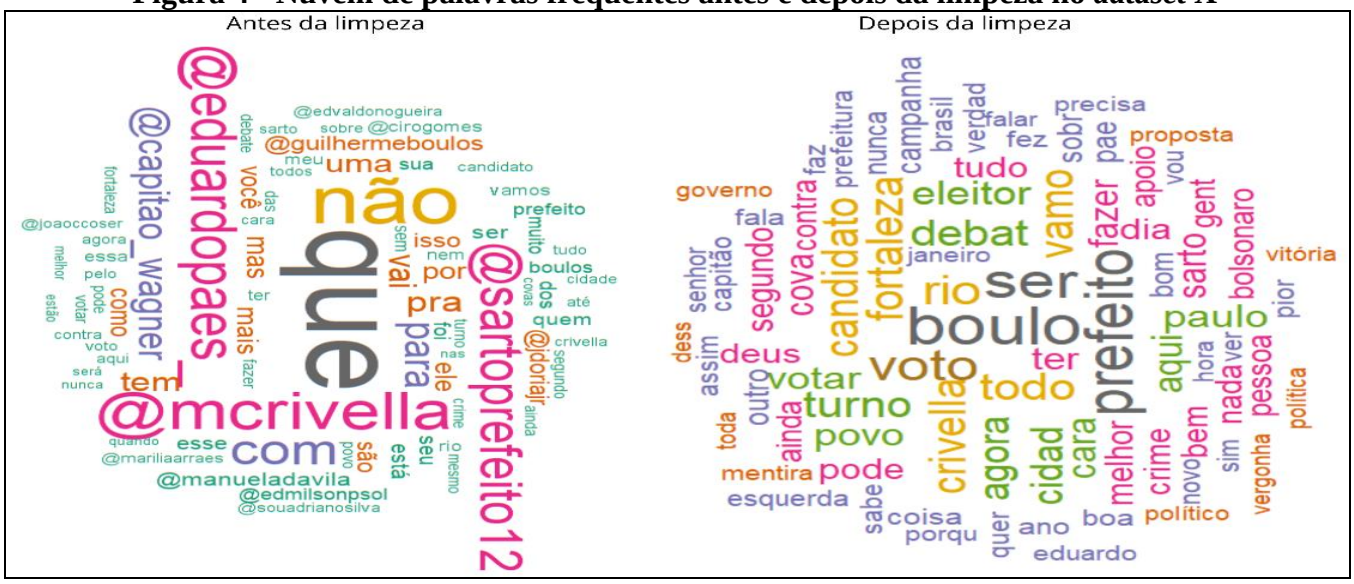
Para limpeza dos dados foram realizadas as seguintes alterações nos *datasets* Kaggle e X: remoção de *stop words*; *steeming*; remoção de pontuações, números e espaços em branco excedentes; remoção de caracteres *hashtags* e *links*; e conversão para letras minúsculas. O pacote *SnowballC* (BOUCHET-VALAT, 2023) da linguagem R foi utilizado para realizar a limpeza. As Figuras 3 e 4 apresentam as nuvens de palavras extraídas dos *datasets* Kaggle e X antes e depois da etapa de limpeza dos dados. Como pode-se verificar, antes da limpeza é alta a frequência de caracteres e palavras sem valor semântico (*stop words*).

Figura 3 - Nuvem de palavras frequentes antes e depois da limpeza no *dataset* Kaggle



Fonte: Elaboração própria.

Figura 4 - Nuvem de palavras frequentes antes e depois da limpeza no *dataset* X



Fonte: Elaboração própria.

RESULTADOS E DISCUSSÃO

No processo CRISP-DM, as etapas 4 (modelagem) e 5 (avaliação) envolvem resultados obtidos, seja durante a construção, ou avaliação do modelo. Sendo assim, a apresentação destas etapas é centralizada na presente seção.

A etapa de modelagem envolveu a codificação dos dados, a construção e o treinamento do classificador *Naïve Bayes*. Conforme previamente descrito, neste trabalho os dados (publicações) utilizados para treinamento do classificador foram codificadas através de uma matriz de frequência de termos criada a partir do *dataset* Kaggle.

O classificador *Naïve Bayes* foi instanciado e treinado em linguagem R utilizando o pacote *e1071* (MEYER et al., 2023). Para evitar superadaptação, foi utilizada a técnica de validação cruzada *k-fold* com $k = 3$. A Tabela 2 apresenta o desempenho do classificador em cada *fold* utilizado no treinamento. Pode-se verificar que o classificador obteve acurácia média de 72,38% e precisão média de 70,73% ao indicar as publicações do *dataset* Kaggle que contém discurso de ódio. O melhor desempenho foi no classificador treinado com o *fold* 1, sendo este o classificador utilizado para avaliação no presente trabalho.

Em comparação aos trabalhos relacionados (estado da arte), Ribeiro (2016) obteve precisão de 97% em seu classificador *Naïve Bayes*. Porém, cabe destacar que foi utilizado um *dataset* com publicações em inglês, enquanto este trabalho utilizou um *dataset* em português. Além disso, o *dataset* utilizado por Ribeiro (2016) foi elaborado com o auxílio de humanos especialistas que

apontaram a polaridade das publicações, e o objetivo do trabalho foi identificar a presença de sarcasmo. Já no *dataset* utilizado para treinamento neste trabalho a classificação foi realizada sem auxílio de humanos (AUGUSTOP, 2018). Quanto ao trabalho de Salminen et al. (2020), o melhor desempenho reportado foi de 92%. Entretanto, foram consideradas publicações de diferentes redes sociais, enquanto o presente trabalho focou exclusivamente na rede social X. Ainda com relação ao estado da arte, Vargas et al. (2022), que também propuseram um classificador *Naïve Bayes*, reportam desempenho parecido ao obtido no presente estudo, com assertividade de 78%. De forma similar, Oliveira et al. (2023) reportaram que o desempenho do ChatGPT para identificação de discurso de ódio ficou em 74%, apesar de sua popularidade enquanto técnica de inteligência artificial.

Tabela 2 - Desempenho do classificador Naïve Bayes após o treinamento

Fold	Acurácia	*Precision*	*Recall*	*F1-score*
1	72,55%	70,77%	70,97%	70,87%
2	72,15%	70,60%	70,16%	70,38%
3	72,45%	70,83%	70,25%	70,53%
Média	72,38%	70,73%	70,46%	70,74%

Fonte: Elaboração própria.

Ao se considerar o estado da arte internacional, cabe ressaltar que o presente trabalho vai de encontro ao que é reportado na literatura, em particular com relação a identificação de discurso de ódio no idioma português do Brasil. Relembrando Tontodimamma *et al*. (2021), o Brasil está entre os 10 países com maior quantidade de pesquisas sobre o debate entre discurso de ódio e liberdade de expressão. Contudo, Yin e Zubiaga (2021), corroborados por Alkomah e Ma (2022), contrapõem com a identificação, na literatura, de poucos conjuntos de dados em idioma português do Brasil passíveis de uso para construção de classificadores e técnicas de aprendizagem de máquina. Similarmente, Jahan e Oussalah (2023) reportam falta de estudos e experimentos com idiomas diferentes do inglês. Neste sentido, fica evidente a relevância do presente trabalho ao considerar publicações no idioma português do Brasil durante a construção e avaliação do modelo preditivo que busca identificar discurso de ódio.

Para realizar a avaliação (etapa 5 da Figura 2), o classificador *Naïve Bayes* foi aplicado no *dataset* X de publicações de candidatos dos municípios selecionados. A saída do classificador indicou a presença ou não de discurso de ódio nas publicações. A referência considerada para avaliar o desempenho foi obtida submetendo o mesmo *dataset* à *Perspective* API para identificação da presença de discurso de ódio.

A *Perspective* API é um projeto de pesquisa colaborativo, disponibilizada gratuitamente para fins não comerciais, que explora o aprendizado de máquina como uma ferramenta para reduzir a toxicidade em discussões online (JIGSAW, 2022). O projeto é resultado de uma pesquisa da equipe de

tecnologia antiabuso do *Google*. A equipe lança regularmente conjuntos de dados, pesquisas acadêmicas e código aberto como parte de seu compromisso com a transparência e a inovação no processamento de linguagem natural e aprendizado de máquina. A *Perspective* API tem sido utilizada por renomadas instituições de publicação de conteúdo (tais como o *The New York Times*, *El Pais*, e *Reddit*) para moderar comentários e participações de seus leitores.

Ao receber uma sentença, a *Perspective* API determina um escore de toxicidade. A toxicidade reflete uma sentença rude, desrespeitosa ou irracional que provavelmente fará o leitor abandonar a discussão e o engajamento na publicação (JIGSAW, 2022). De acordo com Salminen *et al*. (2020), a toxicidade é um indicativo de discurso de ódio. Da mesma forma, Almerekhi *et al*. (2019) empregaram o termo toxicidade como sinônimo de discurso de ódio. Sendo assim, a toxicidade apontada pela *Perspective* API é utilizada como referência para avaliar o desempenho do classificador *Naïve Bayes* proposto neste trabalho.

O escore de toxicidade é um percentual que indica a probabilidade da sentença ser interpretada como tóxica pelo seu leitor (JIGSAW, 2022). Por exemplo, o escore de toxicidade determinado pela *Perspective* API para a sentença *"você é um idiota"* é 0,8, indicando 80% de probabilidade de ser interpretada como tóxica. Para ser considerado como discurso de ódio, este trabalho utiliza o valor 0,7 como limiar do escore de toxicidade. Sendo assim, sentenças com escore maior ou igual este valor são consideradas como discurso de ódio na comparação da *Perspective* API com o classificador *Naïve Bayes* proposto. Este valor de limiar foi adotado com base em estudo de caso da *Perspective* API realizado com uma grande plataforma de jogos online da Europa (JIGSAW, 2019).

A Tabela 3 apresenta os resultados obtidos com o classificador *Naïve Bayes* e com a *Perspective* API na identificação de discurso de ódio nas publicações do *dataset* X. O primeiro aspecto evidenciado por estes resultados é que, independentemente do classificador utilizado, ao menos uma ocorrência de discurso de ódio foi identificada para todos os candidatos.

Tabela 3 - Identificação de discurso de ódio em publicações na rede social X

Município	Candidato	Qtd.	*Naïve Bayes*		*Perspective* API	
		Posts	Qtd.	%	Qtd.	%
São Paulo	Bruno Covas	1209	271	22,42%	241	19,93%
	Guilherme Boulos	5407	1175	21,73%	848	15,68%
Rio de Janeiro	Eduardo Paes	2187	482	22,04%	438	20,02%
	Marcelo Crivella	3231	609	18,85%	667	20,64%
Fortaleza	Cap. Wagner	1654	253	15,29%	293	17,71%
	Sarto	2545	440	17,29%	208	8,17%
Joinville	Adriano Silva	405	180	44,44%	208	51,23%
	Darci de Matos	33	4	12,12%	5	15,15%
Blumenau	João P. Kleinubing	99	4	4,04%	2	2,02%
	Mário Hildebrandt	82	2	2,24%	4	4,87%

Fonte: Elaboração própria.

Nos três municípios mais populosos, verifica-se que o percentual de publicações onde os classificadores identificaram discurso de ódio oscila entre 8,17% (Sarto/*Perspective* API) e 22,42% (Bruno Covas/*Naïve Bayes*). Nos municípios de São Paulo e Rio de Janeiro, os percentuais com discurso de ódio são similares para ambos os candidatos e classificadores. Nestes dois municípios, a maior diferença entre os candidatos é em São Paulo com o classificador *Naïve Bayes* (4,25%). Isso evidencia que nestes municípios o discurso de ódio se manifestou com similar intensidade independente de viés político. Já no município de Fortaleza, o classificador *Naïve Bayes* identificou percentuais similares para os candidatos, enquanto a *Perspective* API apontou uma diferença de 9,54% nas publicações com discurso de ódio entre os candidatos.

Já nos municípios catarinenses, os percentuais de publicações onde foi identificado discurso de ódio são diferentes dos três municípios mais populosos. Em Joinville é onde foi identificado o maior percentual de publicações com discurso de ódio (proporcionalmente à quantidade de publicações) pelos dois classificadores. Por outro lado, em Blumenau estão os menores percentuais.

Ao comparar o classificador *Naïve Bayes* com a *Perspective* API, nota-se desempenho similar na maioria dos candidatos. As maiores diferenças são notadas nos candidatos Sarto, Adriano Silva, e Guilherme Boulos, para os quais o classificador *Naïve Bayes* identificou um percentual diferente de publicações com discurso de ódio em relação à *Perspective* API (9,12%, 6,79% e 6,05% respectivamente). As demais diferenças entre o classificador *Naïve Bayes* e a *Perspective* API ficaram abaixo de 3,1%. Isto indica um desempenho equivalente entre os classificadores, evidenciando que o classificador *Naïve Bayes* desenvolvido atende o objetivo de identificar discurso de ódio em publicações na rede social X e, portanto, tem potencial para uso prático.

Posteriormente à coleta de dados secundários do repositório Kaggle[4] que subsidiaram o treinamento do classificador apresentado neste trabalho, novos *datasets* foram disponibilizados na literatura (VARGAS *et al*., 2022; TRAJANO *et al*., 2023). Estes novos *datasets* podem ser utilizados em trabalhos futuros, visando aperfeiçoar o desempenho do classificador *Naïve Bayes* proposto. Além disso, a disponibilidade destes *datasets* por outros autores posteriormente a realização do presente trabalho reforçam a relevância da identificação de discurso de ódio em redes sociais na língua portuguesa, em particular no Brasil, conforme foi realizado no presente trabalho.

CONSIDERAÇÕES FINAIS

Este estudo descreveu a construção e avaliação de um classificador *Naïve Bayes* para identificar discurso de ódio em publicações na rede social X destinadas a candidatos que disputaram o segundo

turno das eleições municipais de 2020 no Brasil. Para treinamento do classificador foi utilizado um conjunto de publicações da rede social X disponível no repositório Kaggle. Para avaliação, foram coletadas as publicações na rede social X relacionadas aos candidatos de cinco cidades onde houve segundo turno.

Os resultados obtidos na etapa de treinamento do classificador *Naïve Bayes* proposto indicaram que sua acurácia média na identificação de discurso de ódio em publicações é de 72,38%. Este resultado está alinhado com o estado da arte em trabalhos que também utilizaram o classificador *Naïve Bayes* com publicações em idioma português do Brasil (78%), bem como com resultados reportados para o ChatGPT (74%) quando aplicado na tarefa de identificar discurso de ódio.

Na avaliação com publicações da rede social X relacionadas aos candidatos que disputaram o segundo turno das eleições municipais de 2020, o classificador proposto obteve resultado similar à *Perspective API*, uma ferramenta utilizada por instituições de publicação de conteúdo e plataformas de jogos para identificar de discurso de ódio e que foi adotada como referência no presente estudo. A diferença de desempenho entre o classificador *Naïve Bayes* proposto e a *Perspective API* ficou abaixo de 9,12% em todos os candidatos considerados, evidenciando o potencial de aplicação do classificador em publicações no idioma português do Brasil.

A identificação de discurso de ódio de forma automatizada é apontada pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO, 2021) como uma alternativa à identificação manual através de moderadores humanos. Isso se deve ao fato que a tarefa de identificar o discurso de ódio, quando realizada por um moderador humano, é bastante intensiva, consome tempo e custos significativos, e não escala para atender o grande volume de conteúdo produzido em redes sociais atualmente. Neste sentido, a UNESCO destaca iniciativas que empregam técnicas de aprendizagem de máquina para identificar o discurso de ódio. Contudo, enfatiza que muitas dessas iniciativas estão limitadas ao idioma inglês. Isso corrobora a importância de iniciativas como a do presente trabalho, que aplica aprendizagem de máquina para identificar discurso de ódio em publicações no idioma português do Brasil.

Recentemente, o grupo de trabalho do Ministério dos Direitos Humanos e da Cidadania apresentou relatório de recomendações para enfrentamento ao discurso de ódio e ao extremismo no Brasil (DUNKER *et al*., 2024). As recomendações incluem judicializar e responsabilizar os divulgadores e fiadores do ódio, aperfeiçoar bases de dados e análise de informações, bem como a formação de rede de inteligência entre órgãos de segurança pública, sociedade civil e universidades. Cabe destacar que para responsabilizar, o primeiro passo é identificar os divulgadores do discurso de ódio. Neste sentido, o presente estudo vem de encontro a essas recomendações, pois apresenta um

modelo preditivo capaz de identificar discurso de ódio, construído por meio de trabalho acadêmico universitário.

As sugestões de trabalhos futuros incluem analisar estratégias para melhorar o desempenho do classificador *Naïve Bayes*. Isto pode requerer que especialistas em discurso de ódio elaborem um conjunto de dados especificamente relacionado ao tema. Outra sugestão inclui avaliar o classificador proposto em outras eleições ou contextos nos quais a identificação de discurso de ódio possa ser importante, como por exemplo em comentários de notícias ou grupos em aplicativos de conversas.

REFERÊNCIAS

ALKOMAH, F.; MA, X. "A Literature Review of Textual Hate Speech Detection Methods and Datasets". **Information**, vol. 13, n. 6, 2022.

ALMEREKHI, H. *et al*. "Detecting toxicity triggers in online discussions". **30th ACM Conference on Hypertext and Social Media**. New York: ACM, 2019.

AUGUSTOP. "Portuguese Tweets for Sentiment Analysis". **Kaggle** [2018]. Disponível em: <www.kaggle.com>. Acesso em: 28/06/2024.

BISPO, F. "Polícia investiga racismo e ameaça de morte contra vereadora eleita em Joinville". **Estadão** [2020]. Disponível em: <www.estadao.com.br>. Acesso em: 05/06/2024.

BOUCHET-VALAT, M. "Package SnowballC". **The Comprehensive R Archive Network** [2023]. Disponível em: <www.cran.r-project.org>. Acesso em: 28/06/2024.

BRASIL. **Constituição da República Federativa do Brasil**. Brasília: Planalto, 1988. Disponível em: <www.planalto.gov.br>. Acesso em: 19/05/2024.

BRASIL. **Projeto de Lei n. 7582**. Brasília: Planalto, 2014. Disponível em: <www.planalto.gov.br>. Acesso em: 19/05/2024.

CARVALHO, A. C. P. L. F. *et al*. **Ciência de Dados**: Fundamentos e Aplicações. Rio de Janeiro: Grupo GEN, 2024.

DAVIDSON, T. *et al*. "Automated Hate Speech Detection and the Problem of Offensive Language". **International AAAI Conference on Web and Social Media**. Montreal: AAAI, 2017.

DUNKER, C. I. L. *et al*. **Relatório de Recomendações para o Enfrentamento do Discurso de Ódio e o Extremismo no Brasil**. Brasília: Ministério dos Direitos Humanos e da Cidadania, 2024.

ELIAS, M. O.; BRASIL, P. Z. S. "O papel das cortes constitucionais no enfrentamento aos ataques e na defesa da democracia". **Boletim de Conjuntura (BOCA)**, vol. 17, n. 50, 2024.

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science**: Algoritmos de machine learning e métodos de análise. São Paulo: Editora Casa do Código, 2020.

FACELI, K. *et al*. **Inteligência Artificial**: Uma abordagem de aprendizado de máquina. Rio de Janeiro: Editora LTC, 2021.

GARGARELLA, R. "Constitucionalismo y libertad de expresión". *In*: ORDOÑEZ, M. P. A. *et al*. (eds.). **Libertad de expresión**: debates, alcances y nueva agenda. Quito: Unesco, 2011.

GO, A.; BHAYANI, R.; HUANG, L. "Twitter sentiment classification using distant supervision". **Conference Information Systems and Technologies**. Palo Alto: CS224N Project, 2009.

IBGE - Instituto Brasileiro de Geografia E Estatística. **Cidades e Estados do Brasil**. Rio de Janeiro: IBGE, 2022. Disponível em: <www.ibge.gov.br>. Acesso em: 05/01/2022.

JAHAN, M. S.; OUSSALAH, M. "A systematic review of hate speech automatic detection using natural language processing". **Neurocomputing**, vol. 546, n. 1, 2023.

JIGSAW. "One of Europe's largest gaming platforms is tackling toxicity with machine learning". **Medium** [2019]. Disponível em: <www.medium.com>. Acesso em: 28/06/2024.

JIGSAW. **Using machine learning to reduce toxicity online**. New York: JIGSAW, 2022. Disponível em: <www.perspectiveapi.com>. Acesso em: 28/06/2024.

MEYER, D. *et al*. **Misc Functions of the Department of Statistics, Probability Theory Group**. **The Comprehensive R Archive Network** [2023]. Disponível em: <www.cran.r-project.org>. Acesso em: 28/06/2024.

OLIVEIRA, A. S. *et al*. "How Good Is ChatGPT For Detecting Hate Speech In Portuguese?". **Anais do 14º Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre: SBC, 2023.

ONU - Organização das Nações Unidas. **United Nations Strategy and Plan of Action on Hate Speech**. New York: ONU, 2019. Disponível em: <www.un.org>. Acesso em: 28/06/2024.

PEREIRA, J. R. G.; MEDEIROS, O. R.; COUTINHO, C. S. "Regulação do discurso de ódio: análise comparada em países do Sul Global". **Revista de Direito Internacional**, vol. 17, n. 1, 2020.

PROVOST, F.; FAWCETT, T. **Data Science para Negócios**. Rio de Janeiro: Editora Alta Books, 2016.

RAWAT, T.; "Applying CRISP-DM Methodology in Developing Machine Learning Model for Credit Risk Prediction". **Lecture Notes in Networks and Systems**. vol. 739, n. 1, 2023.

RIBEIRO, D. A. **Avaliação do desempenho em métodos de análise de sentimentos e no algoritmo Naïve Bayes** (Trabalho de Conclusão de Curso de Graduação em Sistemas de Informação). Marabá: Unifesspa, 2016.

SAFERNET. "Crimes de ódio têm crescimento de até 650% no primeiro semestre de 2022". **Safernet** [2022]. Disponível em: <www.safernet.org.br>. Acesso em: 28/06/2024.

SALMINEN, J. O. *et al*. "Developing an online hate classifier for multiple social media platforms". **Human-centric Computing and Information Sciences**, vol. 10, n. 1, 2020.

SILVA, N. F. F. **Análise de sentimentos em textos curtos provenientes de redes sociais** (Tese de Doutorado em Ciência da Computação e Matemática Computacional). São Carlos: USP, 2016.

SILVA, V. R. "Eleições de 2018 têm pico de denúncias de discurso de ódio, apontam dados da Safernet". **Associação Gênero e Número** [2018]. Disponível em: <www.generonumero.media>. Acesso em: 28/06/2024.

TONTODIMAMMA, A. *et al*. "Thirty years of research into hate speech: topics of interest and their evolution". **Scientometrics**, vol. 126, n. 1, 2021.

UNESCO - United Nations Educational, Scientific and Cultural Organization. **Addressing hate speech on social media: contemporary challenges**. Paris: Unesco, 2021.

VARGAS, F. *et al*. "HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection". **Thirteenth Language Resources and Evaluation Conference**. Marseille: European Language Resources Association, 2022.

WIRTH, R.; HIPP, J. "CRISP-DM: Towards a standard process model for data mining". **4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining**. Manchester: Practical Application Company, 2000.

X CORP. **About the X API**. San Francisco: X Corp, 2024. Disponível em: <www.developer.x.com>. Acesso em: 28/06/2024.

YIN, W.; ZUBIAGA, A. "Towards generalisable hate speech detection: a review on obstacles and solutions". **PeerJ Computer Science**, vol. 7, 2021.

ZAVALETA-SÁNCHEZ, E. *et al*. "Comparative Study of KDD and CRISP-DM Methodologies for Phishing Identification". **Ninth International Congress on Information and Communication Technology**. London: Springer, 2024.

BOLETIM DE CONJUNTURA (BOCA)

Ano VI | Volume 19 | Nº 55 | Boa Vista |2024

http://www.ioles.com.br/boca