UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLOGIA DA INFORMAÇÃO
CURSO DE SISTEMAS DE INFORMAÇÃO

Leandro Freire Cabral

Detecção de discurso de ódio transfóbico na rede social X mediante técnicas de aprendizado de máquina

ANGICOS
2024

Leandro Freire Cabral

Detecção de discurso de ódio transfóbico na rede social X mediante técnicas de aprendizado de máquina

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Adriana Mara Guimarães de Farias, Prof.ª Ms.

ANGICOS

2024

C117d Cabral, Leandro.
Detecção de discurso de ódio transfóbico na rede social X mediante técnicas de aprendizado de máquina / Leandro Cabral. - 2024.
47 f. : il.

Orientadora: Adriana Farias.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Sistemas de Informação, 2024.

1. transfobia. 2. aprendizado de máquina. 3. discurso de ódio. 4. twitter. I. Farias, Adriana , orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automáto em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Angicos
Bibliotecário: Helder Romero Maia Duarte
CRB: 15/673

Leandro Freire Cabral

Detecção de discurso de ódio transfóbico na rede social X mediante técnicas de aprendizado de máquina

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Sistemas de Informação.

Defendida em: 23/10/2024

BANCA EXAMINADORA

_____
Adriana Mara Guimarães de Farias, Prof.ª Ms (UFERSA)
Presidente

_____
Araken de Medeiros Santos, Prof. Dr. (UFERSA)
Membro Examinador

_____
Ramiro de Vasconcelos dos Santos Junior, Prof. Dr (UFERSA)
Membro Examinador

RESUMO

A transfobia é uma realidade cruel no Brasil, com o país liderando o ranking de assassinatos de pessoas trans por 15 anos consecutivos. As redes sociais, apesar de ferramentas de comunicação e expressão, também são palco para a disseminação de ódio contra a comunidade trans. O discurso de ódio transfóbico é um problema prevalente em plataformas online, como o X, antigo *Twitter*. Este trabalho propõe a criação e avaliação de modelos de aprendizado de máquina para a detecção automática de discurso de ódio transfóbico em textos em português no contexto brasileiro. Os algoritmos de aprendizado de máquina utilizados serão: *Support Vector Machine*, *Random Forest*, *Passive Aggressive Classifier*, *Gaussian Naïve Bayes* e *Multi-Layer Perceptron*. Este trabalho visa contribuir para o combate à transfobia no ambiente online, oferecendo uma ferramenta para a identificação e classificação de discurso de ódio transfóbico. A pesquisa também visa ampliar o conhecimento sobre as características do discurso de ódio transfóbico no contexto brasileiro.

Palavras-chave: transfobia; aprendizado de máquina; discurso de ódio; *Twitter*.

ABSTRACT

Transphobia is a harsh reality in Brazil, with the country leading the ranking of murders of trans people for 15 consecutive years. Social media, despite being tools of communication and expression, are also platforms for the dissemination of hatred against the trans community. Transphobic hate speech is a prevalent problem on online platforms, such as X, formerly known as Twitter. This study proposes the creation and evaluation of machine learning models for the automatic detection of transphobic hate speech in Portuguese texts in the Brazilian context. The machine learning algorithms used will be: Support Vector Machine, Random Forest, Passive Aggressive Classifier, Gaussian Naïve Bayes and Multi-Layer Perceptron. This work aims to contribute to combating transphobia in the online environment by providing a tool for the identification and classification of hate speech. The research also aims to broaden knowledge about the characteristics of transphobic hate speech in the Brazilian context.

Keywords: transphobia; machine learning; hate speech; Twitter.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

OMS	Organização Mundial da Saúde
CID	Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde
STF	Supremo Tribunal Federal
LGBTQIA+	Lésbicas, Gays, Bissexuais, Transexuais, Queer, Intersexuais, Assexuais e demais orientações sexuais e identidades de gênero
PLN	Processamento de Linguagem Natural
SVM	Support Vector Machine
RF	Random Forest
GNB	Gaussian Naive Bayes
MLP	Multi-Layer Perceptron
PAC	Passive Aggressive Classifier
TF-IDF	Term Frequency - Inverse Document Frequency
URL	Uniform Resource Locator
SGD	Stochastic Gradient Descent
NLTK	Natural Language Toolkit
LR	Logistic Regression
XGB	XGBoost
BoW	Bag of Words

SUMÁRIO

1 INTRODUÇÃO

Redes sociais são um tipo de mídia extremamente popular na atualidade, sendo consumidas por diferentes faixas etárias e classes sociais, sendo o Brasil eleito o terceiro país que mais consome redes sociais no mundo no ano de 2022 (BRASIL É O 3º PAÍS QUE MAIS USA REDES SOCIAIS NO MUNDO, 2023). O grande volume de usuários neste tipo de mídia, faz com que seja difícil analisar com precisão o conteúdo publicado, havendo espaço para a propagação de discurso de ódio, principalmente a transfobia. Segundo Podestá (2019), o termo transfobia é um conceito em ascensão, que designa e analisa as violências cometidas contra pessoas trans, representando um grupo heterogêneo de violências específicas que atingem mulheres transexuais, travestis, homens trans, pessoas não binárias, entre outras.

Por muito tempo a transexualidade foi considerada um transtorno mental, sendo descrita na quarta edição do Manual Diagnóstico e Estatístico de Transtornos Mentais como "desordem de identidade de gênero" (STRYKER, 2017), o que contribuiu para reforçar o estigma social já existente em relação a esta parcela da população. No ano de 2018, a Organização Mundial da Saúde (OMS) publicou a décima primeira edição da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID), que retirou a transexualidade da lista de doenças e distúrbios mentais (MINISTÉRIO DOS DIREITOS HUMANOS E DA CIDADANIA, 2018). Em 2019, o Supremo Tribunal Federal (STF) aprovou a criminalização da homofobia e da transfobia no Brasil, enquadrando os atos como crime de racismo, incluindo as declarações publicadas em meios de comunicação, como publicações em redes sociais, podendo ter uma pena de dois a cinco anos, além de multa (OLIVEIRA; BARBIERI, 2019). A remoção da transsexualidade da lista de distúrbios mentais, juntamente com a criminalização da transfobia, demonstram que houve progressos significativos na luta pelos direitos de pessoas trans, embora ainda haja uma longa jornada a percorrer até que esse tipo de preconceito seja devidamente enfrentado.

Honeycutt e Herring (2009), definem o X, antigo Twitter, como um serviço de microblog, onde os usuários podem mandar mensagens (chamadas de *tweets*) limitadas a 140 caracteres em uma interface *web*, na qual o usuário pode decidir se suas mensagens serão públicas ou privadas. J.K. Rowling, a escritora que criou a saga "Harry Potter", causou repercussão ao fazer uma série de comentários transfóbicos na plataforma X, após o anúncio de uma possível legislação britânica que poderia criminalizar comentários transfóbicos. A escritora chegou até mesmo a afirmar preferir ser presa do que parar com os comentários

(PRADO, 2023). O caso ganhou visibilidade na mídia pela fama da autora, no entanto, há muitos outros casos de transfobia na plataforma que seguem impunes.

Em dezembro de 2023, ocorreram mudanças nas políticas de uso da plataforma X, em relação ao combate à transfobia. A mudança ocorreu em atendimento a uma decisão judicial motivada pelo Ministério Público Federal. Com as novas políticas, práticas transfóbicas, como se referir a um indivíduo usando pronomes que não correspondem à sua identidade de gênero autoidentificada, ou se referir a uma pessoa trans pelo seu nome de registro antes de sua transição, serão vedadas pela plataforma (‘X’, ANTIGO TWITTER, RESTABELECE MEDIDAS DE COMBATE À TRANSFOBIA APÓS DECISÃO JUDICIAL, 2024).

Pesquisadores têm recorrido à ciência de dados para criar modelos de aprendizado de máquina capazes de detectar discurso de ódio em redes sociais, alguns desses trabalhos são descritos na Sessão 5. D. Kelleher e Tierney (2018) definem a ciência de dados como um conjunto de princípios, resoluções de problemas, algoritmos e processos para a extração de padrões não óbvios de grandes conjuntos de dados, com finalidades específicas. Por meio da ciência de dados, é possível analisar padrões em textos e averiguar se o conteúdo pode ser considerado ofensivo, de acordo com critérios pré-estabelecidos.

Uma das principais ferramentas na análise de dados textuais é o Processamento de Linguagem Natural (PLN), que pode ser complexo de se lidar, já que a linguagem natural (linguagem desenvolvida naturalmente por seres humanos) é ambígua, podendo conter aspectos linguísticos como ironia, sarcasmo ou expressões idiomáticas. O fato de dados textuais não serem exatos como dados numéricos, dificulta muito a análise dos mesmos, reforçando a importância de se elaborar modelos de aprendizado de máquina robustos e devidamente preparados.

Redes sociais possibilitam comunicação e expressão em grande escala, mas também possibilitam a propagação de discursos de ódio, sendo a transfobia um deles, no qual a disseminação de discursos transfóbicos ocorre impunemente em plataformas como o X, . Diante dessa problemática, surge a questão: *é possível criar um modelo de classificação que analise postagens da plataforma X, que possa identificar conteúdo transfóbico?* Este trabalho propõe criar, configurar e avaliar modelos que usam algoritmos de aprendizagem de máquina supervisionada para classificar automaticamente um texto como transfóbico ou não transfóbico. Os algoritmos utilizados serão: Support Vector Machine (SVM), Random Forest (RF), Passive Aggressive Classifier (PAC), Gaussian Naïve Bayes (GNB) e Multi-Layer Perceptron (MLP). Os textos analisados serão em português no cenário brasileiro.

Diante de todo o contexto exposto, nas próximas seções, serão detalhados a justificativa na Seção 2, objetivos geral e específicos na Seção 3, revisão bibliográfica na Seção 4, trabalhos relacionados na Seção 5, metodologia na Seção 6, resultados esperados na Seção 7, resultados obtidos na Seção 8, e por fim, considerações finais na Seção 9.

2 JUSTIFICATIVA

O Brasil está entre os países que mais matam pessoas trans. Segundo Benevides (2024) no ano de 2023, houve um aumento de mais de 10% nos casos de assassinatos de pessoas trans em relação a 2022 e manteve o posto como o país que mais assassinou pessoas trans pelo 15º ano consecutivo. Essa estatística reflete o quanto a transfobia está presente na sociedade brasileira, fato esse que também se manifesta de muitas formas no meio virtual em redes sociais. Diversos estudos têm sido realizados no campo de aprendizado de máquina, com foco em analisar dados referentes a mídias sociais, visando identificar padrões com uma finalidade específica, que, no caso deste trabalho, é a identificação do discurso de ódio contra a comunidade LGBTQIA+, em específico à comunidade trans.

Lima (2022), em seu trabalho de conclusão de curso, propôs criar e avaliar experimentalmente, modelos que usam algoritmos de aprendizagem de máquina para classificar um texto como homofóbico ou não homofóbico.

Antunes, Issa, Hoed (2023), assim como Souza, Nakamura, Nakamura (2022) construíram modelos preditivos no intuito de detectar se um Tweet contém conteúdo ofensivo a indivíduos homossexuais ou não.

Santos, Henriques, Guedes (2022) realizaram um estudo semelhante aos dois anteriores, porém optou-se por usar apenas os comentários feitos em uma publicação do ex-deputado Jean Wyllys, no período em que o mesmo anunciou seu exílio.

Ashraf et al. (2022) em seu artigo, disserta sobre a criação de um sistema elaborado para detectar automaticamente comentários homofóbicos/transfóbicos nas redes sociais. A base de dados utilizada na pesquisa consiste em comentários do YouTube.

Barbosa (2021) em seu trabalho de conclusão de curso, elaborou um modelo capaz de detectar mensagens com teor transfóbico no X, utilizando uma base de dados contendo 720 Tweets, os quais possuíam termos de cunho transfóbico.

Redes sociais, em geral, possuem normas e diretrizes referentes ao conteúdo publicado pelos usuários, podendo haver penalidades para aqueles que forem contra as regras impostas, sendo possível até mesmo haver o banimento da conta do usuário. Podem ser aplicadas técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, para detectar automaticamente termos considerados inapropriados para o ambiente das redes sociais (FARZINDAR; INKPEN, 2017). Através dessas técnicas, é possível aplicar uma penalidade ao autor da publicação, mas, também, há a opção de denunciar uma publicação,

que será analisada por uma pessoa destinada a isso. Entretanto, se houver um grande volume de denúncias, a equipe responsável por revisá-las pode não ser suficiente para atender a demanda, reforçando a importância do aprendizado de máquina para resolver este tipo de problema.

Este trabalho se propõe a contribuir para a área de aprendizado de máquina aplicada à identificação de discursos de ódio, focando especificamente na transfobia, que compõe uma série de violências cometidas a mulheres trans, travestis, homens trans e entre outras identidades de gênero que se diferem da binaridade (homens e mulheres cisgêneros). Espera-se que os resultados desta pesquisa possam ajudar no combate a este preconceito disseminado em redes sociais, apoiando ações automatizadas para moderação de conteúdo, dado o cenário alarmante de violência contra pessoas trans no Brasil.

3 OBJETIVOS

3.1. Objetivo geral:

Desenvolver e avaliar um modelo computacional que identifique e classifique discursos de ódio transfóbicos em plataformas de redes sociais, contribuindo para o monitoramento, a análise e a mitigação do discurso de ódio online.

3.2. Objetivos específicos:

1. Investigar e analisar as características dos discursos de ódio Online: examinar as manifestações de discursos transfóbicos na rede social X para identificar características linguísticas, padrões e contextos que distinguem esses discursos de outros tipos de comunicação.
2. Desenvolver e testar modelos computacionais para classificação: criar e avaliar modelos de mineração de dados, incluindo SVM, RF, PAC, GNB e MLP, para a classificação eficaz de textos como transfóbicos ou não transfóbicos.
3. Aprimorar métodos de pré-processamento e extração de características: implementar técnicas de pré-processamento de dados textuais, incluindo a remoção de ruídos (como URLs e menções), e aplicar o método de extração de características TF-IDF, para melhorar a precisão dos modelos de classificação.
4. Avaliar o desempenho e a aplicabilidade dos modelos desenvolvidos para garantir a versatilidade e a aplicabilidade do sistema na detecção de discursos de ódio.

4 TRABALHOS RELACIONADOS

4.1. Identificação Automática de Tweets Homofóbicos em Português

Lima (2022), em seu trabalho de conclusão de curso, propõe criar, configurar e avaliar experimentalmente, modelos que usam algoritmos de aprendizagem de máquina supervisionada para classificar automaticamente um texto como homofóbico ou não homofóbico. A base de dados utilizada no estudo é composta por *tweets* filtrados por termos específicos considerados homofóbicos. A metodologia utilizada no trabalho é composta por: coleta de dados, processamento dos dados coletados, rotulagem necessária para utilizar as técnicas de aprendizagem de máquina supervisionada, e a construção e avaliação dos modelos. Os dados foram coletados utilizando o *SNScrape*, um *scraper* (ferramenta de software que automatiza a coleta de dados na web) feito em *Python*, especializado em redes sociais. Com essa ferramenta, foi possível coletar o conteúdo textual do tweet, o idioma, uma variável que indica se o tweet é um *retweet* ou não e a descrição da biografia (bio) do usuário que o criou. A busca pelos tweets foi realizada por meio de doze termos considerados homofóbicos, limitando a data de publicação ao dia 07/10/2018.

Antes da rotulagem manual, foi feita uma filtragem dos tweets para agilizar o processo, utilizando um modelo encontrado no *GitHub*, que apresentava boas métricas de desempenho de classificação de discurso de ódio em dados textuais. Após a classificação, 1.713 dados foram rotulados como contendo discurso de ódio ou não, e separados para compor outra base de dados que servirá para treino e teste.

Após a rotulagem, o próximo passo foi a criação dos modelos, utilizando a *Scikit-learn*, uma biblioteca em *Python* especializada em inteligência artificial. Foram utilizadas 14 configurações de algoritmos para a construção dos modelos. Os dados foram utilizados em 25% para teste e 75% para treino. Os dados do conjunto de treinamento foram usados para treinar 14 variações de algoritmos de aprendizagem de máquina. Entre os 14 modelos, foi utilizado o modelo com a melhor *F1-Score*, para classificar os dados como homofóbicos ou não, e utilizá-los como conjunto de treinamento de mais modelos. Ao todo foram criados 70 modelos com 14 variações de algoritmos, comparados conforme os valores de acurácia, *recall*, precisão e *F1-Score*. Analisando o conjunto total de modelos, concluiu-se que as métricas atingidas foram satisfatórias, com 19 dos 70 modelos possuindo *F1-score* alto, que corresponde a 70% ou mais.

4.2. *Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM*

Ashraf *et al*. (2022) em seu artigo, disserta sobre a criação de um sistema elaborado para detectar automaticamente comentários homofóbicos/transfóbicos nas redes sociais. A base de dados utilizada na pesquisa consiste em comentários do *YouTube*, que estão em inglês, tâmil ou contém tanto inglês quanto tâmil. A base de dados possui texto homofóbico e não homofóbico, extraído entre 08/2020 e 02/2021, sendo em três, sendo uma para cada situação, respectivamente.

A metodologia deste trabalho consiste em: pré-processamento textual, extração de características e aplicação dos algoritmos de classificação. O conteúdo textual foi tratado, removendo caracteres especiais, *URLs*, espaços em branco e caracteres repetidos. Após o pré-processamento, foi realizada a extração de características dos comentários utilizando a técnica *Term Frequency- Inverse Document Frequency* (TF-IDF), que representa os comentários como um vetor de números reais. Foram usados cinco algoritmos de classificação neste estudo, sendo eles: SVM, RF, PAC, GNB e MLP. O *F1-Score* foi usado para avaliar todos os modelos.

De todos os classificadores, a base de dados com textos em tâmil foi o que apresentou os maiores valores de *F1-Score*, sendo de 0,70 a 0,85. Já os piores resultados foram de textos em inglês, com 0,37 a 0,43 de *F1-Score*. O classificador *SVM* superou todos os outros classificadores, com o *F1-Score* ponderado (W *F1-score*) apresentando valores altos para todos os modelos (de 0,88 a 0,92), enquanto o macro *F1-score* (M *F1-score*) apresentou os menores valores (de 0,39 a 0,84). O classificador SVM obteve uma acurácia mínima de 0,90.

4.3. Técnicas de Machine Learning Aplicadas a Mineração de Dados e Análise de Sentimentos Para Predição de Homofobia no Twitter

Antunes, Issa, Hoed (2023) construíram um modelo preditivo no intuito de detectar se um Tweet contém conteúdo ofensivo a indivíduos da comunidade LGBTQIA+ ou não. A base de dados utilizado na pesquisa é composta por aproximadamente 3000 *tweets* de conteúdo diverso, podendo ser comentários ofensivos à comunidade LGBTQIA+ ou não. Para a análise dos dados, foi utilizada a linguagem *Python* e o ambiente de desenvolvimento *Google Colab*.

A metodologia utilizada foi a *CRISP-DM*, composta por seis etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem dos dados, avaliação e implementação. Para o entendimento do negócio, o principal objetivo foi a identificação e

denúncia de textos e publicações homofóbicas feitas por usuários. Para o entendimento dos dados, se iniciou o processo de mineração de dados, onde foram selecionados os termos considerados homofóbicos, usados como palavras-chave para realizar uma busca por tweets utilizando uma *API*.

Na etapa de preparação de dados foi realizada a limpeza e formatação dos tweets para depois classificá-los como zero para não ofensivo, e um para ofensivo. A modelagem foi realizada utilizando a *API Scikit-learn*, que fornece a classe *SGDClassifier* para implementar o método de Descida de Gradiente Estocástica (*SGD*), que serve para a construção de um estimador para problemas de classificação. Foi utilizada a função *train test split* para dividir os dados de aprendizado de máquina, sendo 70% para conjunto de treinamento e 30% para conjunto de testes. Na validação, Antunes, Issa e Hoed (2023) avaliam os modelos para garantir que eles atendem às necessidades do negócio e na implantação são exibidos os resultados.

Na exibição dos resultados foi utilizada uma matriz de confusão e a partir dela foram analisadas as métricas de precisão, *recall*, *F1-score* e acurácia. O modelo obteve 93% de precisão na classe zero (não ofensivo). Já a classe um (ofensivo) obteve 62% de precisão, sendo destacado o ocorrido devido ao desbalanceamento da base de dados. O algoritmo obteve 86% de acurácia nos resultados e mesmo com a baixa precisão da classe um, a alta precisão da classe zero cumpre com o objetivo destacado, o qual é de diminuir ao máximo os falsos positivos. Ao final do artigo, é ressaltada a pretensão de, futuramente, coletar mais dados para balancear a base de dados.

4.4. Detecção de Discurso de Ódio: Homofobia

Souza, Nakamura, Nakamura (2022) elaboraram um modelo de aprendizagem de máquina capaz de identificar se uma mensagem possui teor homofóbico ou não. Utilizou-se um conjunto de dados retirado do Twitter, que contém informações sobre discursos homofóbicos.

A metodologia do trabalho ocorreu nas seguintes etapas: coleta de dados, análise e filtragem de dados, rotulagem dos dados, construção de características e treinamento do modelo de classificação.

A coleta de dados ocorreu extraindo mensagens do Twitter via *API Rest*. Foram utilizados os termos "*fag*", "*faggot*" e "*bigotry*" como palavras-chave para realizar a busca pelos tweets na *API*, limitando-os à língua inglesa. Em seguida, os tweets obtidos foram

rotulados manualmente, por meio de um site de autoria própria. Os usuários podem acessar o site após realizarem um cadastro com informações como gênero, sexualidade e tempo de uso diário do twitter, que serão considerados posteriormente na análise dos dados. Feito o cadastro, os usuários podem visualizar os tweets e opinar se os mesmos possuem teor homofóbico ou não.

Para a construção de características, utilizou-se a biblioteca *Python* chamada *Natural Language Toolkit* (*NLTK*), para calcular a análise de sentimento sobre cada tweet. Os tweets foram filtrados para remover *stopwords*, palavras de frequência alta no texto e que não apresentam relevância para o contexto principal, como pronomes, conjunções, preposições, etc. Foram retirados dos textos dos tweets n-gramas de tamanho de um a três caracteres, além da frequência das palavras de cada tweet. Para isso foi utilizado o *Term Frequency-Inverse Dense Frequency* (*TF-IDF*) que consiste em uma técnica utilizada para determinar um peso para a semântica de sentenças. Essa técnica foi aplicada para cada palavra de um tweet e para cada carácter de cada palavra de um tweet.

Para o treinamento do modelo de classificação foram utilizados os algoritmos SVM, *XGBoost* (XGB), *Logistic Regression* (LR) e RF.

O modelo RF obteve o melhor desempenho, com precisão de 0,72, revocação de 0,93 e F1-Score de 0,81 na classificação de tweets homofóbicos, enquanto na classificação de tweets não-homofóbicos, obteve-se uma precisão de 0,98, revocação de 0,91 e *F1-Score* de 0,94. Ao final do artigo, é ressaltada a pretensão de realizar futuramente a inclusão de mais classes, referentes a outros tipos de preconceito além da homofobia.

4.5. Reconhecimento de Mensagens com Teor Transfóbico no Twitter

Barbosa (2021) em seu trabalho de conclusão de curso, aplicou estratégias de aprendizado de máquina para detectar mensagens com teor transfóbico no twitter. Foi utilizada uma base de dados contendo 720 *tweets*, os quais possuíam termos de cunho transfóbico.

Para a metodologia deste trabalho, foram realizadas as seguintes etapas: seleção de expressões de cunho transfóbico, coleta de tweets através das palavras-chave, pré-processamento dos dados e classificação.

As expressões transfóbicas foram selecionadas para serem utilizadas como palavras-chave na busca por publicações no X. Os termos foram escolhidos em março de

2021, mediante artigos e veículos e notícias consideradas transfóbicas. Em seguida, foi utilizada a API do *Twitter* para coletar 1096 *tweets*.

Na etapa de pré-processamento, foram removidos: menções, links, emoticons e símbolos de pontuação mediante funções e expressões regulares. Além disso, houve um processo de remoção de *stopwords* (palavras de frequência alta no texto e que não apresentam relevância para o contexto principal), usando a biblioteca *Natural Language Toolkit* (NLTK). Foi utilizado o método *BoW* (*Bag of Words*), que consiste em separar o texto e palavras desconsiderando as palavras repetidas, e, em seguida, criar vetores binários que serão usados no algoritmo.

Para a classificação dos tweets foram utilizados o *Multinomial Naive Bayes*, a regressão logística e algoritmos de *boosting,* que consistem em utilizar um conjunto de modelos simples que trabalharão em conjunto, resultando em um desempenho superior e minimizando erros de treinamento. Os algoritmos utilizados foram: *AdaBoost* e *Gradient Bosting*, todos provindos da biblioteca *Scikit-learn*. No processo de treinamento do modelo, a base de dados foi dividida em 70% para treino e 30% para teste. Para treinar os modelos utilizados, foi utilizado o método de validação cruzada *GridSearchCV* através das métricas de acurácia, precisão, *recall* e *F1-Score*, também provindas da biblioteca *Scikit-learn*. Os quatro modelos utilizados atingiram desempenhos razoáveis ou ótimos, com valores de acurácia, precisão, *recall* e *F1-Score* entre 65,33% e 84,72%. Dentre os modelos, o que mais se destacou foi o *Gradient Boosting*. Foram consideradas melhorias futuras nos modelos pensando no potencial evolutivo do trabalho e os resultados satisfatórios obtidos.

4.6. O Discurso de Ódio Homofóbico no Twitter a partir da Análise de Dados

Santos, Henriques, Guedes (2022) realizaram um estudo visando averiguar, por análise de dados, mensagens do Twitter com conteúdo homofóbico. A base de dados utilizada contém 2597 tweets, que consistem em comentários feitos no perfil do ex-deputado Jean Wyllys no período em que o mesmo anunciou seu exílio.

A metodologia deste trabalho é composta pelas seguintes etapas: coleta dos dados, pré-processamento, transformação dos dados e criação de modelos de agrupamento. Os dados foram coletados a partir da *API* do twitter, onde a busca foi realizada através do *ID* da publicação na qual o referido ex-deputado informa sobre seu exílio. Foi utilizada a ferramenta *Hatebase,* a qual é um repositório online de discurso de ódio multilíngue. A base de dados foi filtrada com base nos termos homofóbicos registrados no *Hatebase*.

O texto de todos os tweets passou por um pré-processamento, que inclui a remoção de conteúdos como *URLs*, emojis e imagens, remoção de *stopwords*, redução de todas as palavras ao seu radical e a conversão dos textos para minúsculo. Foi utilizado o método *Term Frequency- Inverse Document Frequency* TF-IDF para estimar a importância de cada termo no texto. Em seguida, se iniciou a etapa de mineração de dados, onde se utilizou nuvens de palavras para identificar os termos mais frequentes nos textos, juntamente com a análise de bigramas, para verificar a incidência das mesmas para o entendimento da homofobia.

O estudo se propôs a identificar como os textos podem ser agrupados em termos similares para poder apresentar uma possível evidência de como os termos homofóbicos podem ser separados, podendo fomentar estudos futuros. A publicação do ex-deputado Jean Wyllys conta com 2587 respostas, das quais foram feitas por 827 usuários únicos, estimando cerca de duas mensagens por pessoa contendo termos do *Hatebase*. A nuvem de palavras utilizada no estudo aponta os seguintes termos como sendo mais frequentes: “viado”, “bicha”, “biba” e “gay”. Na análise de bigramas, ficou evidente que a palavra “bicha” teve muita relação com as palavras “burra” e “enrustida”, o que ressalta como a ofensa se refere às capacidades racionais do sujeito ou a maneira como o mesmo se reconhece enquanto sujeito. A análise de agrupamento adotada no estudo foi o método do Cotovelo, que indicou o melhor número de clusters sendo cinco. Dos grupos definidos, nota-se que alguns dos termos se misturam, sendo possível que seja necessário realizar uma segunda abordagem para avaliar o quão boa foi a decisão. No cluster um, o termo “biba” é o mais frequente, com 20% de representação em relação a todos os termos do cluster. Para o cluster dois, o termo “viado” se destaca com uma frequência de 33%. No cluster três, o termo “viado” também se destaca, mas com um percentual de 13% de frequência, que também se destaca no cluster quatro, mas com 9,6% de frequência. Por fim, no cluster cinco, se evidenciam termos que por si só não são homofóbicos, mas que se relacionam com o discurso homofóbico presente nos tweets, que foi evidenciado através dos bigramas, como o termo “enrustido”.

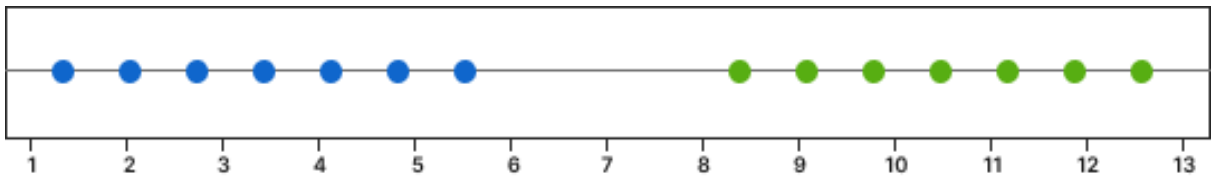
5 REFERENCIAL TEÓRICO

5.1. Support Vector Machine (SVM)

Support Vector Machine, ou máquina de vetores de suporte, é um modelo de aprendizado de máquina supervisionado, que pode ser utilizada em tarefas de classificação e regressão, sendo capaz de performar classificação linear ou não linear, e até mesmo detectar *outliers*, valores atípicos em uma base de dados (GÉRON, 2019). O objetivo da SVM é encontrar um hiperplano que maximize a margem entre os dados de treinamento e o próprio hiperplano. Essa margem é a distância entre os pontos de cada classe mais próximos do hiperplano (chamados de vetores de suporte).

Na Figura 1.1, pode-se observar representação de um conjunto de instâncias de um atributo de uma base dados, separado em duas classes:

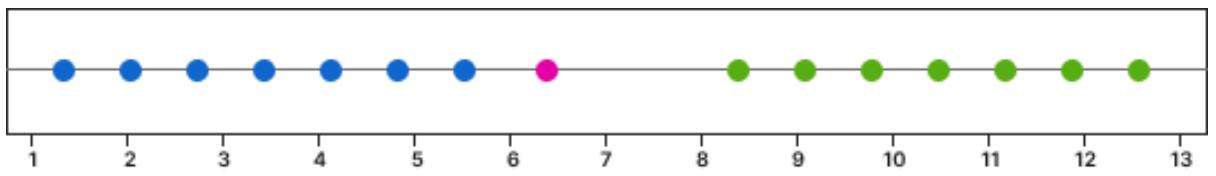
Figura 1.1 – Exemplo de separação de classes



Fonte: Elaboração própria.

Se for inserido mais um ponto, como mostrado na Figura 1.2, pode-se assumir que esse ponto pertencerá ao grupo azul, ou que pertenceria ao grupo verde, caso estivesse mais a direita.

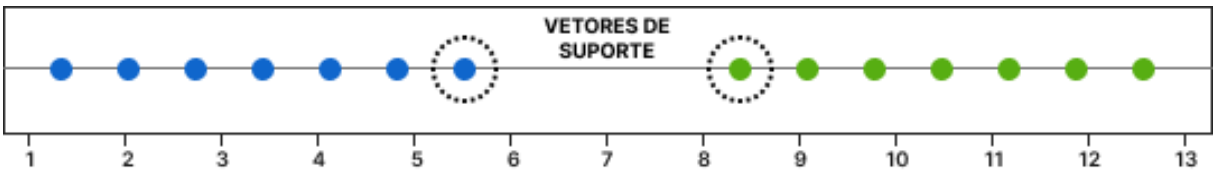
Figura 1.2 – Exemplo de separação de classes



Fonte: Elaboração própria.

Essa lógica se assemelha à estratégia do SVM, onde os pontos mais próximos entre duas classes serão utilizados para traçar um hiperplano, que será responsável por separar as classes. Os dois pontos mais próximos entre as duas classes são os vetores de suporte, como mostra a Figura 1.3.

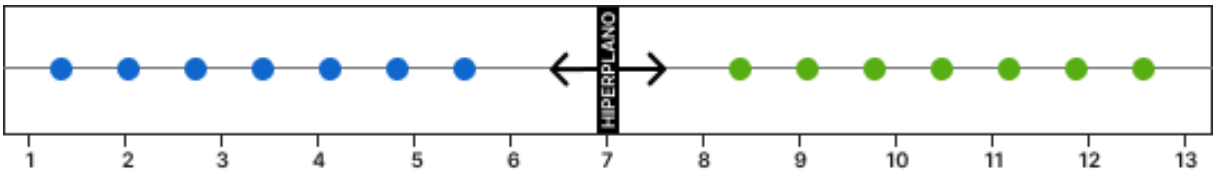
Figura 1.3 – Vetores de suporte



Fonte: Elaboração própria.

O hiperplano se trata da margem que separa ambas as classes, como mostra a Figura 1.4:

Figura 1.4 – Hiperplano



Fonte: Elaboração própria.

5.2. Random Forest (RF)

Uma *Random Forest*, ou floresta aleatória, é um algoritmo de aprendizado de máquina do tipo *ensemble learning*, ou seja, combina vários modelos para obter um resultado mais preciso e robusto, onde os modelos combinados são árvores de decisão. As árvores de decisão são algoritmos de aprendizado de máquina, definidos por Almeida, Carvalho, Menino (2017) como formas de representação de conhecimento via uma estrutura hierárquica de perguntas na forma *if-then-else* (controle de fluxo comumente usado em linguagens de programação). Nessa estrutura existem nós, utilizados para representar perguntas, e desses, se ramificam outros nós, que podem representar uma resposta ou outra pergunta.

5.3. Passive Aggressive Classifier (PAC)

Passive Agressive Classifier, ou classificador passivo agressivo, é uma família de algoritmos do tipo online-learning, ou seja, os dados são inseridos sequencialmente e o modelo de aprendizado de máquina é atualizado a cada inserção (PASSIVE AGGRESSIVE CLASSIFIERS, 2023). Estes algoritmos são chamados de passivo agressivo devido como funcionam, onde no caso de uma predição correta, o algoritmo reagirá passivamente, não ocorrendo nenhuma mudança no modelo, no entanto, caso haja uma predição incorreta, a reação será agressiva, ocorrendo mudanças no modelo.

5.4. Gaussian Naïve Bayes (GNB)

O Gaussian Naïve Bayes é um algoritmo de aprendizado de máquina supervisionado comumente utilizado para tarefas de classificação. Ele é baseado no teorema de Bayes e assume fortemente a independência condicional entre as *features* (características) para prever a classe de um novo dado. Por exemplo, se uma fruta for vermelha, redonda e com 10 cm de diâmetro, pode ser considerada uma maçã. Independente de relações entre as características, um classificador Bayesiano assume que cada uma destas características contribui de forma independente para a probabilidade de esta fruta ser uma maçã, por esse motivo é chamado de *naive*, que, do inglês, significa ingênuo (SABRY, 2023).

5.5. Multi-Layer Perceptron (MLP)

O *Perceptron* é uma das arquiteturas mais simples de Redes Neurais Artificiais (RNA). É baseado em um neurônio artificial chamado de *threshold logic unit* (TLU), ou unidade lógica de limiar, na qual as entradas e a saída são números, e cada conexão de entrada está associada a um peso. O TLU calcula uma soma ponderada de suas entradas, depois aplica uma função a essa soma e gera o resultado (GÉRON, 2019). O MLP, ou Perceptron Multicamadas, é composto por uma camada de entrada, uma ou mais camadas de TLU, e uma camada final de TLU específica para as saídas (GÉRON, 2019). MLP é comumente usado em classificação de textos e análise de sentimentos.

5.6. Term Frequency- Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF), é uma técnica estatística usada para avaliar a importância de uma palavra em um documento em um *corpus* (conjunto de textos ou documentos usados para análise linguística e processamento de linguagem natural). O objetivo desta técnica é ponderar a frequência de uma palavra em um documento pelo seu peso em todo o corpus.

Term Frequency (TF) representa a frequência com que uma palavra aparece em um texto. TF mede a frequência de uma palavra em um documento. Manning, Raghavan e Schütze (2008) explicam que a forma mais simples de calcular o TF é atribuir ao peso de uma palavra o número de vezes que ela ocorre no documento:

Figura 2.1 – *Term Frequency*

$$\mathrm{TF}(t, d) = \frac{\text{Número de vezes que o termo } t \text{ aparece no documento } d}{\text{Número total de termos no documento } d}$$

Fonte: Elaboração própria.

Inverse Document Frequency (IDF), é uma métrica que ajusta a importância de um termo em um corpus de documentos, ajudando a destacar termos que são mais específicos para alguns documentos e não tão comuns em outros. Manning, Raghavan e Schütze (2008) apontam que, quando se analisa somente a frequência bruta dos termos, todos os termos são considerados igualmente relevantes, o que se torna um problema crítico, já que, dependendo do contexto, a frequência de um termo pode não estar diretamente ligada à sua relevância. O conceito por trás do IDF é que termos que aparecem em muitos documentos são menos informativos para discriminar documentos relevantes. Por outro lado, termos que aparecem em poucos documentos têm um peso maior porque são mais raros e, portanto, mais relevantes para distinguir entre documentos. O IDF soluciona esse problema, balanceando a frequência dos termos e ajustando sua importância, considerando a rareza do termo em todo o corpus. O IDF é calculado como o logaritmo do número total de documentos dividido pelo número de documentos que contêm a palavra:

Figura 2.2 – *Inverse Document Frequency*

$$\mathrm{IDF}(t, D) = \log\left(\frac{\text{Número total de documentos } D}{\text{Número total de termos no documento } d}\right)$$

Fonte: Elaboração própria.

Combinando ambas as definições, obtém-se um peso composto para cada termo em cada documento (MANNING; RAGHAVAN; SCHÜTZE, 2008). A fórmula completa do TF-IDF é:

Figura 2.3 – Term Frquency - *Inverse Document Frequency*

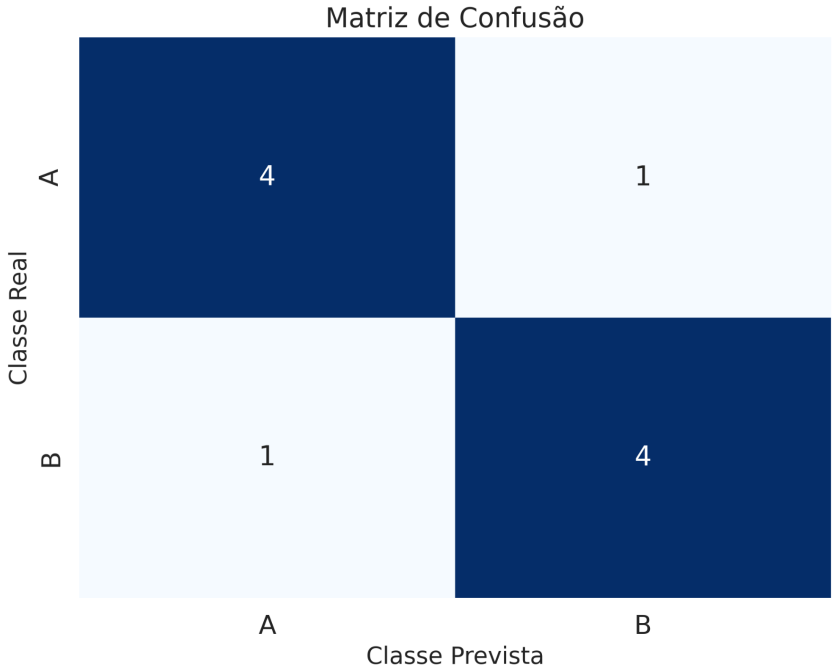
$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D)$$

Fonte: Elaboração própria.

5.7. Matriz de confusão

A matriz de confusão é uma maneira de avaliar o desempenho de um modelo de classificação. Para Géron (2019), a ideia geral de uma matriz de confusão é contar quantas vezes uma classe A foi classificada como B, ou seja, quantas predições incorretas o modelo realizou. Na Figura 3.1, em azul estão as predições corretas, enquanto as predições erradas estão em branco.

Figura 3.1 – Exemplo de matriz de confusão



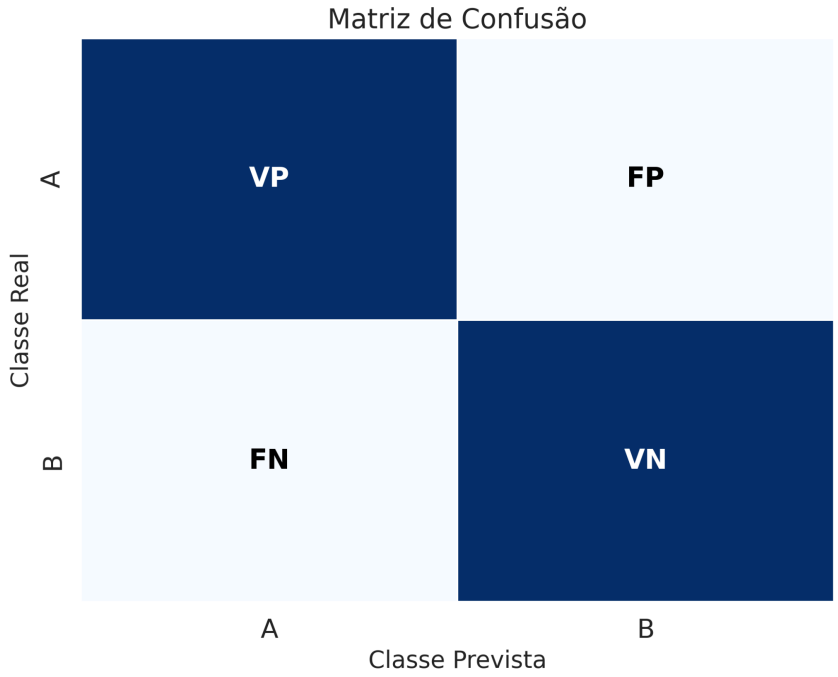
Fonte: Elaboração própria.

Na Figura 3.1, 4 instâncias foram preditas corretamente como pertencentes à classe A, sendo classificadas como verdadeiros positivos (VP). Uma instância foi predita como pertencente à classe B, quando, na verdade, era da classe A, caracterizando um falso negativo (FN). Além disso, uma instância foi predita como pertencente à classe A, quando, na verdade, era da classe B, caracterizando um falso positivo (FP). Por fim, 4 instâncias foram preditas corretamente como pertencentes à classe B, sendo classificadas como verdadeiros negativos (VN). A Figura 3.2 exemplifica a localização das seguintes classificações:

- **Verdadeiro Positivo (VP):** O modelo previu corretamente que uma instância pertence a uma classe específica. Por exemplo, quando o modelo classifica um caso positivo como positivo.
- **Falso Negativo (FN):** O modelo não conseguiu identificar uma instância que realmente pertence a uma classe, prevendo-a como não pertencente. Isso ocorre quando o modelo classifica um caso positivo como negativo.
- **Falso Positivo (FP):** O modelo incorretamente classifica uma instância como pertencente a uma classe, quando, na verdade, não pertence. Isso acontece quando o modelo identifica um caso negativo como positivo.

- **Verdadeiro Negativo (VN):** O modelo previu corretamente que uma instância não pertence a uma classe específica. Isso ocorre quando o modelo classifica um caso negativo como negativo.

Figura 3.2 – Exemplo de matriz de confusão



Fonte: Elaboração própria.

5.8. Métricas de avaliação dos modelos

Acurácia

Reflete a proporção de previsões corretas (tanto positivas quanto negativas) em relação ao total de previsões, oferecendo uma visão geral do desempenho.

Precision (Precisão)

Para Géron (2019), a precisão do classificador se trata da acurácia das predições corretas entre todas as instâncias previstas como uma determinada classe. Por exemplo, se o modelo prevê “Transfóbico”, a precisão reflete quantas dessas previsões pertencem realmente a essa classe. A fórmula da precisão se dá por:

Figura 4.1 – Precisão

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Fonte: Elaboração própria.

VP é o número de verdadeiros positivos e FP é o número de falsos positivos.

Recall (Revocação)

Para Géron (2019), a revocação, também chamada de sensibilidade ou taxa de verdadeiros positivos, é a taxa de instâncias positivas preditas corretamente pelo classificador. Esta métrica mede a capacidade do modelo de identificar corretamente todas as instâncias de uma determinada classe entre todas as instâncias verdadeiras dessa classe. Alta revocação significa que o modelo consegue capturar a maioria dos exemplos reais da classe. Sua fórmula se dá por:

Figura 4.2 – Revocação

$$\text{Revocação} = \frac{VP}{VP + FN}$$

Fonte: Elaboração própria.

FN é o número de falsos negativos.

F1-Score

Géron (2019) descreve o F1-score como a média harmônica entre precisão e revocação, que aumenta o peso de valores menores, ao contrário de uma média comum, que trata todos os valores igualmente. Sua fórmula é:

Figura 4.3 – *F1-Score*

$$F_1 = \frac{2}{\frac{1}{\text{Precisão}} + \frac{1}{\text{Revocação}}} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} = \frac{2}{VP + \frac{FN + FP}{2}}$$

Fonte: Elaboração própria.

6 METODOLOGIA

A metodologia de pesquisa deste trabalho é definida como uma abordagem quantitativa e computacional. Foram aplicadas técnicas de mineração de dados para analisar e classificar discurso de ódio transfóbico na rede social X. As etapas desta metodologia envolvem a coleta, o pré-processamento e a análise dos dados textuais, seguida pelo desenvolvimento, teste e avaliação de modelos de classificação.

O aspecto quantitativo desta pesquisa se dá por utilizar dados numéricos extraídos de textos e os analisar por meio de métodos estatísticos e algoritmos de mineração de dados, fornecendo uma abordagem quantificável para a identificação de discursos de ódio. Já o caráter computacional da pesquisa, se dá por usar ferramentas computacionais e algoritmos para processar e analisar os dados, dependendo de algoritmos especializados e a linguagem de programação *Python*.

Foram aplicadas técnicas de processamento de linguagem natural para tratar e extrair características dos dados textuais, possibilitando a análise semântica e sintática dos textos. Modelos de classificação baseados em aprendizado supervisionado foram treinados com conjuntos de dados rotulados para identificar automaticamente discursos de ódio.

Esta metodologia destina-se não apenas a identificar e classificar discursos de ódio com precisão, mas também a contribuir para a criação de ambientes digitais mais seguros e inclusivos, oferecendo *insights* para pesquisadores, desenvolvedores de políticas de moderação de conteúdo e educadores.

6.1. Coleta de Dados

Para a coleta de dados, foi utilizada a ScraperAPI, uma API de raspagem de dados, ou *web scraping*, o qual é o processo de extrair informações de sites da web. Para realizar a busca pelos dados, foram reunidos termos considerados transfóbicos, que serviram como palavras-chave. Os termos foram escolhidos considerando expressões amplamente utilizadas em discurso de ódio relatado em notícias. Também foram escolhidos termos relacionados ao debate sobre a causa trans, que não estão necessariamente relacionados ao discurso de ódio, já que a intenção é distinguir textos transfóbicos e não transfóbicos.

Os termos escolhidos foram: “mudança de sexo”, “traveco”, “trans”, “transexuais”, “transexual”, “transexualidade”, “travecos”, “travecão”, “travecões”, “homem vestido de mulher”, “não existe homem trans”, “não existe mulher trans”, “transgênero”, “transgêneros”,

“nasceu mulher”, “nasceu homem”, “homem que se diz mulher”, “mulher que se diz homem”, “quer ser mulher”, “quer ser homem”, “homem vestido de mulher”, “nunca será mulher”, “ideologia de gênero”, “trans esporte”, “travesti”, “travestis”, “comunidade trans”, “transição de gênero”, “afirmação de gênero”, “direitos trans”, “violência trans”, “transfobia no Brasil”, “nunca será homem”.

Na busca, foi utilizado um filtro para que fossem retornados apenas tweets brasileiros. Concluída a busca, a API retorna uma base de dados contendo tweets que correspondem a palavra-chave inserida. A base de dados contém o título do tweet (title), conteúdo do tweet (snippet) e termos em destaque (*highlighs*) como mostra a Figura 5.1. Ao unir todos os dados coletados, a base de dados resultante continha 3221 instâncias.

Figura 5.1 – dados extraídos

	position	title	snippet	highlighs
0	1	O termo "traveco" é pejorativo e só contribui para aumentar ...	Oct 22, 2015 — O termo "traveco" é pejorativo e só contribui para aumentar a violência e discriminação contra as trans e travestis. Translate post.	['traveco']
1	2	"Mesma coisa para o termo "traveco" ...	Dec 1, 2015 — Mesma coisa para o termo "traveco"! O sufixo "eco" pode indicar inferioridade e adicionar qualidades negativas: ambos são pejorativos!	['traveco']
2	3	Por que o termo “traveco” é transfóbico e não deve ser usado	Jan 21, 2022 — BBB: Por que o termo “traveco” é transfóbico e não deve ser usado https://revistaforum.com.br/blogs/popnoticias/por-que-nao-usar-traveco/...	['traveco', 'traveco']
3	4	"traveco é termo pejorativo. travesti é uma identidade de ...	Jan 21, 2022 — traveco é termo pejorativo. travesti é uma identidade de gênero. pois mais que o senso comum tenha medo de falar a palavra travesti e nos ...	['traveco']
4	5	Tracklist	Jan 21, 2022 — Rodrigo conversou com Linn da Quebrada após ele falar o termo pejorativo “traveco” na última noite no #BBB22. Durante a conversa, o brother ...	['traveco']
5	6	Nos Trends BRASIL	Jan 21, 2022 — TRAVECO está #nostrendsbrasil Rodrigo: "Eu tava lembrando da história do pinto do traveco que você contou." Vinicius: "Ei, traveco não. É ...	['TRAVECO', 'traveco', 'traveco']

Fonte: Elaboração própria.

6.2. Rotulagem dos Dados

Uma etapa crucial do treinamento de algoritmos de aprendizado de máquina é rotulagem dos dados. Dos tweets, 1480 foram rotulados manualmente, dos quais foram definidas duas classes: transfóbicos com valor 1 e não transfóbicos com valor 0, sendo 870 classificados como não transfóbicos, enquanto 610 foram classificados como transfóbicos. Devido ao desbalanceamento das classes, é necessário tomar algumas medidas para minimizar o impacto nos resultados.

Para os algoritmos SVM e RF, foi usado o parâmetro class_weight='balanced'. Esse parâmetro ajusta os pesos das classes de maneira inversamente proporcional à sua frequência no conjunto de dados, ou seja, a classe 1 (transfóbico), que está em minoria, recebe um peso maior, enquanto a classe 0 (não transfóbico), que está em maioria, recebe um peso menor, auxiliando o modelo a ser mais sensível às instâncias minoritárias, evitando o favorecimento predominante da classe majoritária.

Os algoritmos Naive Bayes, Multi-Layer Perceptron e Passive Aggressive Classifier não possuem diretamente o parâmetro class_weight, neste caso foi aplicada a técnica SMOTE (Synthetic Minority Over-sampling Technique) para balancear o conjunto de dados. Essa técnica gera novas instâncias sintéticas da classe minoritária, aumentando a representatividade dessa classe no treinamento, o que melhorou o desempenho em termos de precisão e recall para essa classe, que poderia ser sub-representada sem o uso do balanceamento.

6.3. Pré-processamento dos dados

O pré-processamento de dados possui a finalidade de refinar a base de dados, removendo quaisquer elementos que não sejam úteis para a análise. Para isto, foram removidos elementos não textuais, como URLs, emojis, menções e caracteres especiais.
Em seguida, algumas etapas serão realizadas com foco no processamento de linguagem natural:

1. Converter os caracteres em minúsculo, pois os algoritmos podem reconhecer diferenças de capitalização como caracteres distintos.
2. Remoção de stopwords: podem ser artigos, preposições, conjunções, entre outros. Quando se trata de análise de dados textuais, esse tipo de palavra pode ser considerado ruído de dados, já que não carregam significado semântico considerável.
3. Lematização: O objetivo da lematização é reduzir uma palavra à sua forma base ou lema, removendo todas as inflexões, como conjugações ou plurais. Por exemplo, a palavra “correrá” seria convertida para “correr”, sendo a forma infinitiva do verbo. Ao aplicar a lematização, algoritmos de aprendizado de máquina podem tratar diferentes formas de uma palavra (como “correr”, “correu”, “correndo”) como uma única entrada, preservando o sentido semântico comum e evitando que diferentes formas de uma palavra sejam tratadas como termos distintos. Isso melhora a precisão e a eficiência na análise do texto, já que reduz a variação nas palavras que o modelo precisa considerar.

6.4. Extração de Características

A extração de características foi realizada implementando o método TF-IDF para converter o texto em um vetor de características, que destaca a importância relativa das palavras no corpus de dados.

6.5. Desenvolvimento e avaliação dos modelos

Foram desenvolvidos modelos de classificação baseados em aprendizado supervisionado, empregando os algoritmos: SVM, RF, PAC, GNB, e MLP. O conjunto de dados foi dividido em treino (75%) e teste (25%) para avaliar o desempenho dos modelos. Foram utilizadas as métricas de acurácia, precisão, recall e F1-score para avaliar o desempenho de cada modelo.

6.6. Análise dos resultados e implementação

Interpretação dos resultados obtidos para entender a eficácia dos modelos e identificar padrões sobre a natureza do discurso de ódio na plataforma X. A documentação dos modelos desenvolvidos será disponibilizada para a comunidade científica e interessados, incluindo o código-fonte, a metodologia aplicada e os resultados obtidos.

7 RESULTADOS ESPERADOS

Os resultados esperados deste estudo abrangem diversos aspectos relacionados ao desenvolvimento e avaliação de modelos de aprendizado de máquina para a detecção de discurso de ódio transfóbico no X. Espera-se que os modelos treinados alcancem altos níveis de acurácia, *recall* e *F1-Score* ao serem testados em um conjunto de dados contendo texto transfóbico e não transfóbico. É preferível que as métricas citadas atinjam níveis superiores a 80%, superando os modelos descritos em literaturas já existentes.

Por fim, a pesquisa visa fornecer percepções sobre as características e a prevalência do discurso de ódio transfóbico no contexto online brasileiro, esclarecendo sobre o alcance e o impacto desse fenômeno. Ao obter uma compreensão mais profunda das dinâmicas subjacentes ao discurso transfóbico nas redes sociais.

8 RESULTADOS OBTIDOS

No Quadro 1, estão as métricas obtidas em cada modelo:

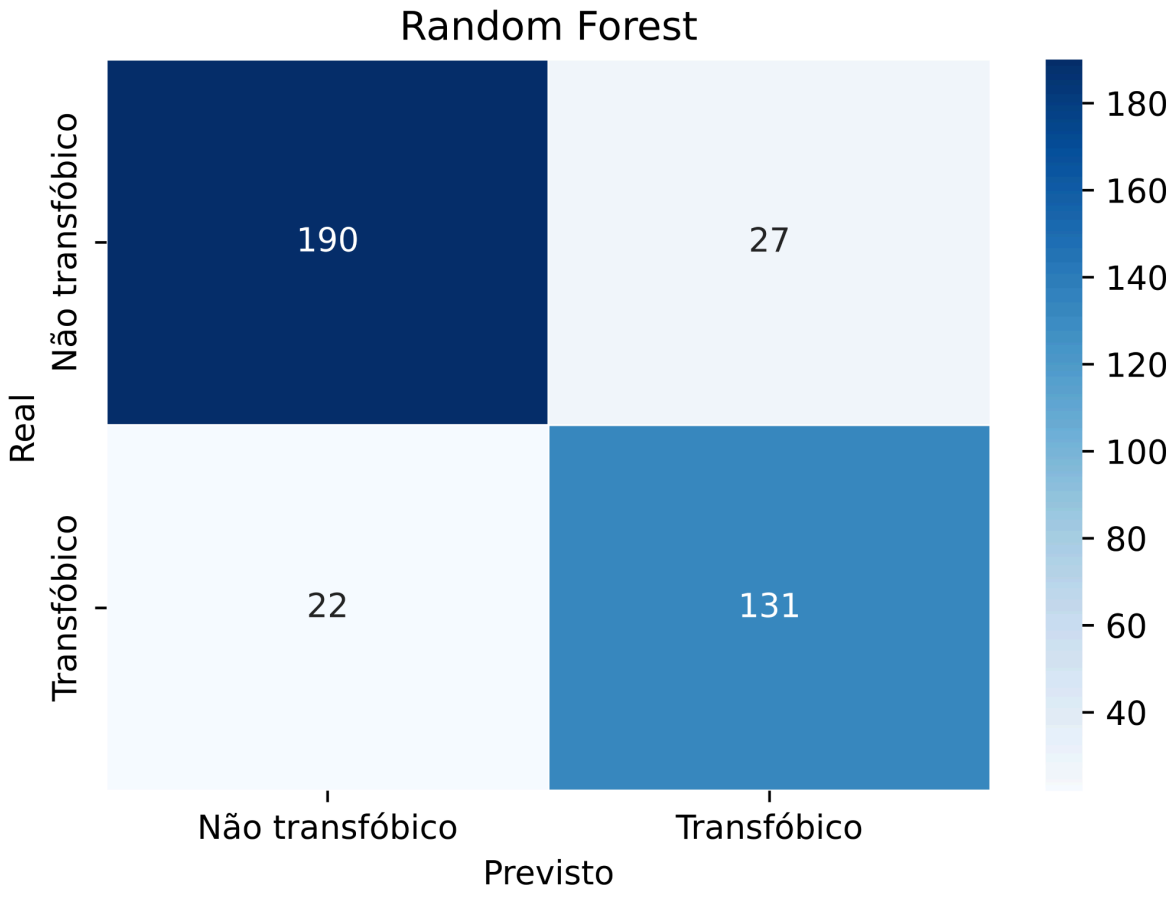
Quadro 1 – Comparação entre modelos

Modelo	Acurácia	Classe	Precision	Recall	F1-Score
Random Forest	86,75%	Transfóbico	83%	86%	84%
		Não transfóbico	90%	88%	89%
Support Vector Machine	84,59%	Transfóbico	79%	86%	82%
		Não transfóbico	89%	84%	86%
Passive Aggressive Classifier	84,05%	Transfóbico	80%	82%	81%
		Não transfóbico	87%	86%	86%
Gaussian Naive Bayes	81,62%	Transfóbico	74%	85%	79%
		Não transfóbico	88%	79%	83%
Multi-Layer Perceptron	81,35%	Transfóbico	80%	74%	77%
		Não transfóbico	82%	87%	84%

O Random Forest foi o modelo de melhor desempenho, apresentando alta precisão e revocação em ambas as classes, isso significa que é um modelo confiável para predição, com baixa probabilidade de erro. Support Vector Machine e Passive Aggressive Classifier tiveram resultados similares, com uma leve tendência a favorecer a classe “Não transfóbico”. Gaussian Naive Bayes e Multi-Layer Perceptron ficaram atrás, com dificuldades para prever corretamente a classe “Transfóbico”.

8.1. Matriz de confusão

Figura 6.1 – Matriz de confusão do modelo com melhor desempenho



Fonte: Elaboração própria.

A matriz de confusão mostra o desempenho do Random Forest, o modelo que se sobressaiu em relação aos outros. Cada célula da matriz contém o número de amostras que caíram em cada combinação de real e previsto:

- **Verdadeiros negativos (190):** O modelo classificou corretamente 190 instâncias como "Não transfóbico", ou seja, eram realmente não transfóbicos previstos corretamente.
- **Falsos positivos (27):** O modelo classificou erroneamente 27 instâncias como "Transfóbico", quando, na verdade, eram "Não transfóbico".

- **Falsos negativos (22):** O modelo classificou erroneamente 22 instâncias como “Não transfóbico”, quando, na verdade, eram “Transfóbico”.
- **Verdadeiros positivos (131):** O modelo corretamente classificou 131 instâncias como “Transfóbico”, ou seja, eram realmente transfóbicos previstos corretamente.
- **Precisão para classe “Não transfóbico”:** A maioria das instâncias não transfóbicas foi corretamente classificada (190 corretas contra 27 incorretas).
- **Precisão para classe “Transfóbico”:** O modelo também foi relativamente eficaz em identificar transfobia, com 131 classificações corretas contra 22 incorretas.

8.2. Nuvem de palavras

Figura 6.2 – Nuvem de palavras da base de dados

Fonte: Elaboração própria.

As palavras que mais se destacaram na base de dados foram: “travesti”, “traveco”, e “mulher”. A presença do termo “traveco” indica a presença de linguagem depreciativa, já que, além de ser um termo pejorativo, a maioria dos comentários que contém essa palavra foram classificados como transfóbicos. Já as palavras “travesti” e “mulher” são usadas predominantemente em um contexto neutro ou autoafirmativo, e não são palavras ofensivas isoladamente.

8.3. Termos mais frequentes em tweets classificados como transfóbicos

Termo	Quantidade de instâncias
mulher	295
traveco	264
homem	242
travecão	109
nascer	44
feminino	31
trans	36
ideologia	44

Dentre os tweets classificados como transfóbicos, as palavras mais recorrentes foram: “mulher”, “traveco”, e “homem”. Isso indica novamente que o discurso transfóbico tende a deslegitimação da identidade de gênero de pessoas trans. Os termos “traveco” e “travecão” aparecem na lista com altas frequências, mostrando que no discurso de ódio, termos explicitamente pejorativos são usados com mais frequência. As palavras “nascer” e “feminino” estão relacionadas à narrativa de que o gênero é imutável e biologicamente determinado.

8.4. Distribuição de classificações entre os termos

Termo	Não transfóbico	Transfóbico
mulher	266	295
homem	169	242
traveco	73	264
trans	189	36
transgênero	197	3
nascer	112	44
travesti	150	11
sexo	120	10
mudança	108	4
travecão	12	109
transsexual	112	2
gênero	61	37
transsexualidade	84	1

É possível notar que termos como “mulher” e “homem” aparecem em abundância em ambos os casos, indicando que, isoladamente, essas palavras não são ofensivas, sendo necessário analisar o contexto. Um dos tweets classificados como transfóbicos diz o seguinte: “*A empresa contratou um homem, que se diz mulher, para fazer propaganda de artigos femininos. Você já sabe o que não deve comprar.*” na declaração em questão, não há nenhum termo explicitamente transfóbico, no entanto, a expressão “homem que se diz mulher” é uma forma de deslegitimar a identidade de gênero de mulheres trans, além de que a publicação em questão incentiva o boicote a propagandas que tenham a presença de pessoas trans.

Os termos “traveco” e “travecão” são quase exclusivamente encontrados em publicações transfóbicas, com uma quantidade significativamente maior na coluna de “Transfóbico” (264 e 109 instâncias, respectivamente). Isso indica que esses termos são amplamente utilizados de maneira pejorativa.

Termos como “transgênero”, “trans”, e “travesti”, aparecem mais frequentemente no contexto não transfóbico (“transgênero” aparecendo 197 vezes no contexto não transfóbico e apenas 3 vezes em contextos transfóbicos). Isso indica que, quando um indivíduo possui a intenção de ofender a comunidade trans, é optado por usar termos explicitamente pejorativos ao invés de termos neutros.

A palavra “nascer” está ligada ao discurso biológico em relação ao gênero, aparece 44 vezes no contexto transfóbico e 112 vezes no não transfóbico. Isso mostra que mesmo termos associados à biologia podem ser apropriados ofensivamente, como no tweet “A mulher mais bonita de Portugal nasceu homem”.

Termos como “mudança” e “gênero” aparecem em ambas as categorias, mas em menor quantidade no contexto transfóbico, o que sugere que esses termos são usados em discussões neutras ou informativas.

9 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo desenvolver e avaliar modelos de aprendizado de máquina para a identificação e classificação de discursos transfóbicos em redes sociais. Com base nos resultados obtidos, o modelo Random Forest apresentou o melhor desempenho, destacando-se em precisão e recall, o que o torna uma ferramenta promissora para a detecção automática de conteúdos transfóbicos. Os modelos Support Vector Machine e Passive Aggressive Classifier, também se mostraram eficazes, mas com uma ligeira tendência a favorecer a classe “não transfóbico”.

A análise revelou que termos como “traveco” e “travecão” são quase exclusivamente utilizados em contextos transfóbicos, enquanto palavras como “transgênero” e “trans” aparecem predominantemente em discursos neutros ou informativos. Isso evidencia que a simples presença de um termo não é suficiente para classificar um texto como ofensivo, sendo o contexto essencial para uma análise precisa.

Esses resultados têm implicações importantes para a moderação de conteúdo em plataformas digitais, no combate à transfobia. Embora o trabalho tenha alcançado resultados promissores, limitações como a dependência do contexto devem ser consideradas para aprimorar os modelos no futuro.

REFERÊNCIAS

ALMEIDA, Adriano; CARVALHO, Felipe; MENINO, Felipe. **Introdução ao Machine Learning**: de alunos para alunos. 2017. Disponível em: https://dataat.github.io/introducao-ao-machine-learning/index.html#nossos-livros-textos. Acesso em: 07 abr. 2024.

ANTUNES, Mayk Brendon Almeida; ISSA, Matheus de Freitas; HOED, Raphael Magalhães. **Técnicas de Machine Learning Aplicada a Mineração de Dados e Análise de Sentimentos para Predição de Homofobia no Twitter**. In: NUEVAS IDEAS EN INFORMÁTICA EDUCATIVA, 16., 2022, Santiago de Chile. Anais […] . Porto Alegre: Jaime Sánchez, 2022. v. 16, p. 331–339. DOI: https://doi.org/10.54751/revistafoco.v16n1-121.

BARBOSA, Iann Carvalho. **Reconhecimento de Mensagens Com Teor Transfóbico no Twitter**. 2021. 14 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal de Campina Grande, Campina Grande, 2021.

Benevides, Bruna G. **Dossiê**: assassinatos e violências contra travestis e transexuais brasileiras em 2023. ANTRA (Associação Nacional de Travestis e Transexuais) – Brasília, DF: Distrito Drag; ANTRA, 2024. 125p.

BRASIL É O 3° PAÍS QUE MAIS USA REDES SOCIAIS NO MUNDO. Brasília, 14 mar. 2023. Disponível em: https://www.poder360.com.br/brasil/brasil-e-o-3o-pais-que-mais-usa-redes-sociais-no-mundo/#:~:text=Levantamento%20da%20Comscore%20mostra%20que,sociais%20(96%2C9%25). Acesso em: 06 fev. 2024.

D.KELLEHER, John; TIERNEY, Brendan. **Data Science**. Cambridge: Mit Press, 2018.

FARZINDAR, Anna Atefeh; INKPEN, Diana. **Natural Language Processing for Social Media**. 2. ed. San Rafael: Morgan & Claypool, 2017.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**: Concepts, Tools, and Techniques to Build Intelligent Systems. 2. ed. Sebastopol: O'Reilly Media, 2019.

HONEYCUTT, Courtenay; HERRING, Susan C. 2009. **Beyond microblogging**: Conversation and collaboration via twitter. pages 1 – 10. Acesso em: 18 fev. 2024.

LIMA, Douglas Pereira de. **Identificação Automática de Tweets Homofóbicos em português**. 2022. 11 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal de Campina Grande, Campina Grande, 2022.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008.

Nsrin Ashraf, Mohamed Taha, Ahmed Abd Elfattah, and Hamada Nayel. 2022. NAYEL @LT-EDI-ACL2022: **Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM**. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 287–290, Dublin, Ireland. Association for Computational Linguistics.

OLIVEIRA, Mariana, BÁRBIERI, Luiz Felipe. **STF permite criminalização da homofobia e da transfobia**. TV Globo e G1, Brasília, 13 de jun. de 2019. Disponível em: https://g1.globo.com/politica/noticia/2019/06/13/stf-permite-criminalizacao-da-homofobia-e-da-transfobia.ghtml. Acesso em: 30 de jan. de 2024.

PASSIVE Aggressive Classifiers. 2023. Disponível em: https://www.geeksforgeeks.org/passive-aggressive-classifiers/. Acesso em: 07 abr. 2024.

PODESTÁ, L. L. de. **Ensaio sobre o conceito de transfobia**. Revista Periódicus, [S. l.], v. 1, n. 11, p. 363–380, 2019. DOI: 10.9771/peri.v1i11.27873. Disponível em: https://periodicos.ufba.br/index.php/revistaperiodicus/article/view/27873. Acesso em: 31 jan. 2024.

PRADO, Pedro Benjamin. **Criadora de "Harry Potter" prefere ser presa a deixar de fazer comentários transfóbicos**. Terra, [s.l.], 20 out. 2023. Disponível em: https://www.terra.com.br/diversao/entre-telas/filmes/criadora-de-harry-potter-prefere-ser-presa-a-deixar-de-fazer-comentarios-transfobicos,2134e13a98d7ff8a245905bbf7f0bafabrq38d6k.html . Acesso em: 06 fev. 2024.

SABRY, Fouad. **Naive Bayes Classifier**: fundamentals and applications. [S.L.]: One Billion Knowledgeable, 2023. 198 p.

SANTOS, Vinícius S. dos; HENRIQUES, Felipe da R.; GUEDES, Gustavo. **O Discurso de Ódio Homofóbico no Twitter a partir da Análise de Dados**. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 11. , 2022, Niterói. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 109-120. ISSN 2595-6094. DOI: https://doi.org/10.5753/brasnam.2022.222916.

SOUZA, Andrey O; NAKAMURA, Eduardo F.; NAKAMURA, Fabíola G. Detecção de **Discurso de Ódio**: Homofobia. In: BRAZILIAN E-SCIENCE WORKSHOP (BRESCI), 16. , 2022, Niterói. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 73-80. ISSN 2763-8774. DOI: https://doi.org/10.5753/bresci.2022.223222.

STRYKER, Susan. (2017). **Transgender History:** The Roots of Today's Revolution. 2ª ed. Seal Press.

THERAPY WISDOM. **Understanding gender dysphoria DSM 4 vs 5**. 2023. Disponível em: https://therapywisdom.com/2023/10/25/understanding-gender-dysphoria-dsm-4vs5/. Acesso em: 20 out. 2024.

‘X’, ANTIGO TWITTER, RESTABELECE MEDIDAS DE COMBATE À TRANSFOBIA APÓS DECISÃO JUDICIAL. São Paulo: Basset Ltda, 02 fev. 2024. Semanal. Disponível em: https://www.cartacapital.com.br/mundo/x-antigo-twitter-restabelece-medidas-de-combate-a-transfobia-a. Acesso em: 05 fev. 2024.

Leandro Freire Cabral

Detecção de discurso de ódio transfóbico na rede social X mediante técnicas de aprendizado de máquina

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Sistemas de Informação.

Defendida em: 23/10/2024

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br ADRIANA MARA GUIMARAES DE FARIAS
Data: 23/10/2024 15:24:58-0300
Verifique em https://validar.iti.gov.br

Adriana Mara Guimarães de Farias, Prof.ª Ms (UFERSA)

Documento assinado digitalmente
gov.br ARAKEN DE MEDEIROS SANTOS
Data: 23/10/2024 11:28:20-0300
Verifique em https://validar.iti.gov.br

Araken de Medeiros Santos, Prof. Dr. (UFERSA)

Documento assinado digitalmente
gov.br RAMIRO DE VASCONCELOS DOS SANTOS JUNIC
Data: 23/10/2024 13:06:34-0300
Verifique em https://validar.iti.gov.br

Ramiro de Vasconcelos dos Santos Junior, Prof. Dr (UFERSA)
Membro Examinador